Authenticity is in the eye of the beholder. Beliefs and perceptions of authentic assessment and the influence on student learning.

Authenticity is in the Eye of the Beholder

Beliefs and perceptions of authentic assessment and the influence on student learning

This research reported here was carried out at the
OpenUniversiteitNederland

In the context of the research school

ICO

(Interuniversity Center for Educational Research)

ISBN 10: 90 358 2411 3
ISBN 13: 978 90358 2411 9

Cover design: Jeroen Storm

Printed by Datawyse, Maastricht, The Netherlands

Authenticity is in the Eye of the Beholder

Beliefs and perceptions of authentic assessment and the impact on student learning

Proefschrift

ter verkrijging van de graad van doctor
aan de Open Universiteit Nederland
op gezag van de rector magnificus
prof. dr. ir. F. Mulder
ten overstaan van een door het
College voor promoties ingestelde commissie
in het openbaar te verdedigen

op vrijdag 10 november 2006 te Heerlen
om 15:30 uur precies

door
Judith Theresia Maria Gulikers

geboren op 5 november 1979 te Maastricht

Promotores
Prof. dr. P. A. Kirschner, Open Universiteit Nederland
Prof. dr. Th. J. Bastiaens, Open Universiteit Nederland

Toegevoegd promotor
Dr. L. Kester, Open Universiteit Nederland

Overige leden beoordelingscommissie:
Prof. dr. R. F. Poell, Universiteit Tilburg
Prof. dr. M. S. R. Segers, Universiteit Leiden
Prof. dr. K. M. Stokking, Universiteit Utrecht
Prof. dr. C. P. M. van der Vleuten, Universiteit Maastricht
Dr. E. Cascallar, Assessment Groups International
Dr. D. H. J. M. Dolmans, Universiteit Maastricht
Dr. R. L. Martens, Universiteit Leiden

Voorwoord

Op deze plek wil ik graag een aantal mensen met naam noemen en bedanken die allemaal een belangrijke rol hebben gespeeld in de afgelopen vier jaar van mijn aio-tijd.

Ten eerste mijn begeleiders, Paul Kirschner, Theo Bastiaens en in het laatste jaar Liesbeth Kester. Bedankt voor jullie vertrouwen in mij vanaf de eerste dag en jullie altijd motiverende woorden. Paul, bedankt voor onze inspirerende discussies en jouw levendige voorbeelden waarmee je mij altijd alle mogelijke perspectieven ergens van liet inzien en bedankt voor de nieuwe uitdagingen die je me steeds hebt gegeven. Theo, door jou ben ik uiteindelijk bij de OTEC terecht gekomen. Bedankt voor de vrijheid en mogelijkheden die je me gegeven hebt, zowel in mijn onderzoek als in mijn nevenactiviteiten. Hierdoor heb ik mij op vele gebieden kunnen ontplooien. En Liesbeth, bedankt dat je meteen bereid en enthousiast was toen ik je in mijn laatste jaar vroeg om mijn begeleider te zijn. Je frisse blik op mijn onderzoek, je kritische vragen die mij dwongen om telkens duidelijk uit te leggen wat ik nu eigenlijk bedoelde en je stimulerende woorden zijn alleen maar positief en inspirerend geweest.

Mijn contactpersonen op de verschillende scholen waar ik mijn onderzoek heb mogen uitvoeren: Maria Pelgrum (Baronie College Gezondheid), Marijke Bijnens (Leeuwenborgh Opleidingen Gezondheid), Frans Bleumer, Lisan van Beurden en Marja van de Broek (Baronie College Sociaal Maatschappelijk Werk). Bedankt voor jullie enthousiasme, de prettige samenwerking en al het geregel dat jullie voor mij hebben gedaan. Naast deze contactpersonen uiteraard heel veel dank aan alle studenten, docenten en praktijkbegeleiders verbonden aan bovengenoemde opleidingen voor hun vrijwillige medewerking.

Hans van Buuren die mij kennis liet maken met AMOS en enthousiast wist te maken voor de Structural Equation Modelling.

Suzan van Dommelen, mijn stagaire, die bijgedragen heeft aan delen van de dataverzameling en analyses. Onze discussie hebben mij steeds gestimuleerd en verder geholpen in mijn denken.

Liesbeth Baartman, mijn assessment-aio-maatje. Dankjewel voor de inhoudelijke discussies en je bereidheid om altijd te helpen. Ik hoop dat we in de toekomst veel contact houden, onze assessment ervaringen blijven uitwisselen en vruchtbare samenwerking tussen onze universiteiten kunnen bewerkstelligen.

Mieke Haemers voor je 'magische oog' waarmee je in no time mijn hele proefschrift doorgeworsteld had.

Desirée Joosten-Ten Brinke en alle OOG-ers van de IPABO. Ik vond het ontzettend leuk en zeer inspirerend om zo intensief met jullie samen te werken aan het in de praktijk brengen van al mijn ideeën over assessment. Het geeft veel motivatie om te zien dat je ideeën worden opgepakt (zij het na heel wat strubbeling) en werken. Ook dankjewel aan alle mensen van de Hogeschool Zuyd-OTEC assessmentgroep en onze collega's van O&O van de Universiteit Maastricht, met wie we geregeld over allerlei assessment problematieken hebben gedebateerd. De discussie tussen onderzoekers en mensen in de praktijk is mijns inziens noodzakelijk en zeer belangrijk om tot goede en nieuwe manieren van assessment te komen.

Alle collega's bij OTEC voor het bieden van een plezierige en uitdagende werkplek, de huidige aio's Amber, Femke, Fleurie, Liesbeth B., Gemma, mijn carpool-maatje, Karen, Pieter, Marieke, Sandra en Wendy en met name de oude garde: Angela, Bas, Dominique, Ellen, Iwan, JW, Liesbeth, Linda, PJ, Ron, Sylvia, Tamara. Als echt indirect Limburg meisje moest ik me tussen jullie bewijzen, maar daardoor heb ik wel ontzettend veel van jullie geleerd, werkgerelateerd, maar vooral ook op persoonlijk gebied.

Dan komen we in de persoonlijke sfeer. Ten eerste, mijn oude vertrouwde vrienden uit Maastricht, met name Annemarie, Astrid, Evelyne, Inge en Martine. Gewoon dankjewel dat jullie er zijn! Een speciaal bedankje is op zijn plaats voor mijn lieve vriendinnetje Esther. Dankjewel voor je oprechte interesse en pogingen om altijd te begrijpen waar ik mee bezig was, je honderden sms-jes tijdens het afschrijven van mijn proefschrift en je positieve, opbeurende woorden als ik even helemaal geen zin meer had. Ik ben erg blij dat ik nu weer dichterbij je kom wonen!

Mijn collega's die ik inmiddels meer vrienden kan noemen: Tamara & Bas, Liesbeth, Linda en PJ. Door alle uurtjes die wij samen door hebben gebracht met lunchwandelingen en etentjes, op congres, in de Efteling of Disneyworld en gewoon in de gangen van chiba, kennen jullie mij inmiddels bijna door en door. Ik weet dat ik bij jullie altijd terecht kan voor wat dan ook en dat voelt goed!

Dan Marjan, André, Pim & Joske en Bas. Ik weet dat ik jullie eigenlijk niet mocht noemen, maar jullie horen inmiddels gewoon bij mijn leven. Dank jullie wel voor jullie altijd oprechte interesse! Er zullen nog veel onderwijskundige discussies volgen.

Pap en mam, Mark & Angela, jullie zijn mijn echte thuis. Pap en mam, ik ben jullie erg dankbaar voor de manier waarop jullie mij in het leven hebben gezet: nuchter, positief en zelfstandig. Ik heb ontdekt dat dit ontzettend belangrijke eigenschappen zijn in de wetenschap en sowieso in "de grote mensen wereld". Mam, zonder jou was ik waarschijnlijk nooit in het onderwijsveld terecht gekomen. Het is leuk om deze passie met jou te delen. Ik ben nog altijd trots als mensen vragen: "Ben jij de dochter van….?" Ik kan alleen maar hopen dat mensen ooit gaan vragen of jij "de moeder van…?" bent.

Dan mijn twee paranimfen, Linda en Tamara. Jullie vertegenwoordigen mijn twee kanten en daarom ben ik vereerd dat jullie op 10 november letterlijk achter mij willen staan. Linda, je bent een echte stimulans. Jouw vriendschap en onze discussies geven mij altijd heel veel energie, motivatie en nieuwe ideeën. Tamara, tja wat zal ik zeggen, eerste indrukken kloppen lang niet altijd. Ondanks dat we totaal verschillend zijn, hebben we aan één oogopslag genoeg. Ik bewonder je gedrevenheid, altijd kritische houding en oprechtheid. Ik heb veel van je geleerd en onze vriendschap is onvoorwaardelijk.

En als laatste natuurlijk Flip. Dankjewel dat je het al zolang met me uithoudt, jij weet dat ik niet altijd lach, en dat je me altijd, vooral nu, de mogelijkheid en vrijheid geeft om me te ontplooien, mijn eigen ding te doen en mijn hart te volgen. Maar dat hart blijft, waar ik ook ben of werk, altijd bij jou.

Contents

Chapter 1

General Introduction

"Authenticity is in the eye of the beholder" means that what one person perceives as being authentic is not necessarily perceived as authentic by someone else. Because of the expected strong influence of student assessment perceptions on student learning, the main aim of the research project described in this thesis is to find out how students perceive assessment authenticity and how this influences their learning. However, since authenticity is both multidimensional and subjective, we first need to gain insight into what facets determine authenticity and then need to determine how these facets are perceived by different beholders, being students as well as other stakeholders in authentic assessment practices (e.g., teachers and practitioners). Moreover, perceptions are influenced by beliefs about authentic assessment and these beliefs are, in turn, based on both previous experiences with professional practice and assessment, and thus, might be different for different beholders. This chapter first describes three developments that made authentic assessment a major issue in competency-based education, followed by a description of the aforementioned variables and their hypothesised relationships. In the end, an overview is given of this thesis.

Three Developments that Lead to Authentic Assessment

The major driving force behind the need for authentic assessment is the perceived gap between what is taught, learnt and assessed in school and what is needed at work (Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004; Boud, 1991). This gap arose from changes in the labour market that resulted in the rise of new kinds of jobs with different requirements for employees. In reaction, education began to change, because students needed to learn different 'things' in a different 'way' to keep up with the changes in the labour market. However, changing student learning requires new methods of assessment as assessment is possibly the most salient variable to influence student learning. These changes typify three developments that led to the need for authentic assessment. The next paragraphs will discuss these three developments in more detail.

The first development is the development of society and economy from a proceduralised, industrial mode to a creative information mode, which is reflected in changes in jobs and job requirements (Birenbaum et al., 2006; Dochy, 2001). In the industrial age, knowledge was rather stable and seen as an objective entity that could be transmitted from one person to the other (Honebein, Duffy, & Fishman, 1993; Sternberg, 1999). Employees required content-specific knowledge and skills that allowed for the performance of routine-based tasks.

Today's information age is characterised by rapidly increasing information, major technological developments, and the globalisation of economies. Knowledge is now seen as a subjective and constantly changing construct that has to be created by every individual anew (Sternberg, 1999). Furthermore, knowledge and learning are seen as context-dependent. Meaningful learning requires the application of knowledge and skills in a realistic context (e.g., Brown, Colllins, & Duguid, 1989; Dall'Alba & Sandberg, 1996). This information society requires knowledge workers (Bereiter, 2002) and competent employees (Birenbaum, 1996) who are independent, creative and flexible problem solvers who can use, apply and develop knowledge and who can acquire the skills needed to deal with new problems within their work. In reaction to these societal changes, education began to change.

The second development reflects the development of education for atomised knowledge and skills acquisition, to education for the development of competencies (Biggs, 1996). Traditional practices that aimed at conveying content knowledge and training of proceduralised, simple skills were characterised by: instruction for knowledge transmission, learning by rote memorisation and drill-and-practice, and assessment that is standardised, objective and knowledge-based or focused on showing routine-based skills (Birenbaum, 2003). Bluntly said, subject matter was transmitted to passive students who had to memorise the information and cough it out in multiple-choice tests or carry out simple skills in a well-defined setting. Moreover, instruction, learning and assessment were treated as separate activities with assessment considered as an afterthought (Biggs, 1996).

Today, knowledge is still important, but as a prerequisite for gaining competence (Miller, 1990; Segers, Dochy, & DeCorte, 1999). Educational practices aiming at the development of competencies can be summarised as follows: instruction that focuses on deep learning and

competency development, learning based on reflective-active knowledge construction, and assessment that is contextualized, interpretative and performance-based (Birenbaum, 2003). Education for competence stimulates students to actively construct their own knowledge, integrate knowledge, skills and attitudes into realistic performances, and detect gaps in their competencies to stimulate future learning. Assessment requires student to show their level of competency development. Moreover, more emphasis is now being placed on integrating these three components, instead of viewing them as separate activities (Birenbaum, 1996; Van Merriënboer, 1997). Biggs (1996) called upon creating a constructive alignment between instruction, learning, and assessment to meet the educational goal. Chapter 2 will explain the idea of constructive alignment and its role in authentic educational practices. In addition, the studies in chapter 6 and 7 take constructive alignment, or student perceptions thereof, into account.

This emphasis on integrating instruction, learning, and assessment, instead of considering assessment as an afterthought, is imperative because of the growing evidence for the strong effect of assessment on instruction and, even more important, on student learning. A striking number of metaphors refer to the strong influence of assessment on student learning such as: *The tail wags the dog* (Gibbs, 1992), *the real test bias* (Frederiksen, 1984), *the washback effect* (Alderson & Wall, 1993), *the backwash effect* (Prodroumou, 1995), the *pre-assessment effect* (Gielen, Dochy, & Dierick, 2003), or *consequential validity* (Messick, 1994). The assessment communicates to students what is deemed important and shows them what they need to know and do in order to succeed in their studies. Assessment has been identified as possibly the single most salient influence on student learning, narrowing students' focus to concentrate only on topics to be examined (i.e., what is to be studied) and shaping their approaches to study (i.e., how it is to be studied) (Ramsden, 1992; Rust, 2002; Scouller & Prosser, 1994).

This leads to the third development, namely the shift from a culture of testing to a culture of assessment (Birenbaum, 1996). Changing assessment practices is inevitable when perspectives on learning and instruction are changing, especially when we consider both the need for constructive alignment and the strong influence of assessment on learning and instruction. Traditional, so-called standardised testing methods representative of the testing culture, such as multiple-choice tests, true/false items or short-answer tests, were increasingly criticised for not being suitable for the changed educational goals aiming at competency development (Brown & Knight, 1994; Fredericksen, 1984; Glaser, 1993; Schuwirth & Van der Vleuten, 2004; Wiggins, 1993). They were considered inadequate for measuring higher-order thinking skills and professional competence and were seen as stimulating students to adopt surface study strategies such as memorisation and reproduction at the expense of deep study activities.

Assessments representative of the assessment culture, on the other hand, are expected to fit with the new educational goals, as they aim at promoting learning and evaluating competency development. They stimulate students to integrate knowledge, skills and attitudes and use them to solve realistic professional tasks (Birenbaum, 1996; Dochy, 2001; Reeves & Okey, 1996). Where traditional tests were atomistic, decontextualized, and focused on summative assessment *of* learning, new assessment are integrated, contextualized, and more focused on formative

assessment *for* learning (Birenbaum & Dochy, 1996; Dochy & McDowell, 1997; Segers, 2003). These characteristics of the assessment culture are more likely to fit with the new educational goal of stimulating students to become competent employees.

How authenticity, the main theme of this thesis, fits in with all of these developments is that authenticity aims at decreasing the gap between the world of the school and the world of work.

The Importance of Authenticity

Authentic instruction and assessment reflect a correspondence between what is learned and assessed and what students are expected to do in the workplace (Boud, 1991; Kerka, 1995). By creating this correspondence, education aims at promoting authentic learning that, in turn, should help students to bridge the gap between learning and working. The idea of authentic learning became popular in learning theories such as situated learning and cognitive apprenticeship (Brown, Collins, & Duguid, 1989; Collins, Brown, & Newmann, 1989) that focus on learning in meaningful contexts (i.e., contexts of work or culture). These theories argue that meaningful, authentic learning requires learning in the working context or at least in the context of everyday life outside of school. Nowadays, even more attention is being paid to smoothening the transition from school to work and authentic learning is argued to be crucial for ensuring that smoother transition (Boshuizen, Bromme, & Gruner, 2004).

In light of creating constructive alignment to stimulate authentic learning, authenticity is important in instruction as well as assessment (Biggs, 1996). This thesis, however, focuses on assessment authenticity, as assessment - as previously stated - is seen by many as the driving force behind learning or changing learning. Authenticity is seen as one of the crucial elements of new modes of assessment representative of the assessment culture (Dochy, 2001; Messick, 1994; Segers, 2003).

Increasing the authenticity of an assessment is expected to be important when the focus of the assessment is on competency development. The reasons for increasing the authenticity of these assessments are twofold. First, to measure, as validly as possible, whether a student is capable of functioning in the world of work. This is the summative side of authentic assessment and deals with construct validity (i.e., does the assessment assess what it aims to assess?). Second, to stimulate students towards deeper learning and the development of professional skills. This is the formative side of authentic assessment and deals with its consequential validity (i.e, what are the effects of the assessment on student learning?). Chapter 2 will explain in more detail why authenticity is expected to be important for both construct validity and consequential validity of competency-based assessments.

Authenticity is thus important for competency assessments. Problematic, however, is that authenticity is a vaguely described concept. Therefore, the following question needs be answered: What exactly is assessment authenticity? In short, this thesis defines assessment authenticity by its resemblance to the professional practice situation. The issue of defining authenticity is thoroughly discussed in chapter 2. This, however, turned out to be complicated by two factors, namely that authenticity is both multidimensional and subjective.

Authenticity as a Multidimensional Concept

Newmann and Wehlage (1993) argued that authenticity is not a dichotomous variable meaning that something is not completely authentic or completely inauthentic. This implies that something, in our case an assessment, can be authentic to a certain degree. In other words, authenticity is a continuum. But this is not all. Authenticity is not simply a continuum, it is multidimensional. The five-dimensional framework, extensively described in chapter 2 and used throughout this thesis, unravels the constituent dimensions of authenticity.

Even if we are successful at unravelling the constituent dimensions of authenticity, we are still faced with the second issue, namely that authenticity is subjective.

The Perception of Authenticity and the Role of Beliefs and Previous Experiences

In assessment practices it is widely recognised that *student perceptions* of assessment characteristics determine what they learn and how they learn (Boud, 1995; Entwistle, 1991; Gibbs, 1999; Gijbels, 2005; Scouller & Prosser, 1994; Scouller, 1995; 1997; 1998; Struyven, Dochy, & Janssen, 2003). This implies that for authentic assessments to stimulate student learning it is imperative that students perceive the assessment as authentic. A problem here is that authenticity is subjective, meaning that what students perceive as an authentic assessment is not necessarily the same as what another person (e.g., a teacher or a practitioner) perceives as an authentic assessment. A major concern is thus to examine what assessment characteristics determine the perception of assessment authenticity, how these characteristics are perceived by different beholders, and how student perceptions thereof influence their learning and professional skill development. These issues are addressed in detail in the following chapters of this thesis.

Furthermore, the subjectivity of authenticity is influenced by beliefs and previous experiences. Whether a person perceives an assessment as being authentic does not only depend on several characteristics inherent to the assessment. Rather the perception of authenticity of a currently encountered assessment depends on the frame of reference a person uses to judge authenticity. A frame of reference that guides the interpretation of new experiences, in turn, is built on previous experiences (e.g., Biggs, 1989; Sternberg, 1999; Van Rossum & Schenk, 1984; Samuelowicz & Bain, 1992). This thesis hypothesises first that what a person perceives as an authentic assessment depends on that person's beliefs about assessment authenticity (i.e., the frame of reference) and second, people build their beliefs about assessment authenticity on their previous experiences with assessments and with professional practice. The reasoning behind these hypotheses will be explained in detail in chapter 4.

If previous experiences with professional practice and with assessments influence a person's beliefs about authentic assessment, then it is reasonable to assume that students, teachers and practitioners differ in their beliefs about authentic assessment, for these groups have all had different experiences. In addition, students with differing degrees or kinds of previous experiences (e.g., as a result of internships) might have different beliefs about what authentic assessment involves. Differences between these groups might become problematic, because beliefs influence not only how new authentic assessments are perceived, they also guide future

behaviour (e.g., Sternberg, 1999; Richardson, 2005; Van Rossum & Schenk, 1984). This means that beliefs about authentic assessment both influence how people (e.g., teachers) develop an authentic assessment as well as how students learn in response to a certain assessment. To develop assessments that are perceived as authentic by students and, as a result, are beneficial for their learning, it is relevant to examine two things, namely (1) whether students differ from other stakeholders in their beliefs about what an authentic assessment involves and (2), if students with different beliefs about authentic assessment perceive different types of assessment as authentic and, as a result, beneficial for their learning. These issues will be addressed in chapter 2, 4, and 7.

From General to Specific: Further Contextualisation of this Thesis

Before giving an overview of the chapters and studies in this thesis, this section describes the educational context in which this research project is conducted, which is the context of Vocational Education and Training (VET) in the Netherlands (in Dutch: Middelbaar BeroepsOnderwijs, abbreviated as MBO). The main issues discussed in this introductory chapter are translated to vocational education practices to give an insight in VET in the Netherlands and to substantiate that these are of major concern in this type of education in which the most fundamental question is: "How can we best prepare youth and adults for the workplace of today?" (Wonacott, 2000, p.1).

Development from the Industrial Mode to the Information Mode: the Rise of the Knowledge Worker

Developing and delivering knowledge workers was initially assumed to only be an issue at the university level, where students are mostly prepared for "thinking" jobs and developing thinking skills was assumed to be the role of academic education (Kerka, 1992). However, jobs at the vocational level have changed drastically with the rise of the information society. They have changed from procedural manufacturing and service jobs, which were prevalent in the industrial age, to jobs for knowledge workers in the information economy (Bereiter, 2002; Dall'Alba & Sandberg, 1996)). However, jobs at the vocational level are mostly well-defined and less broad than the future jobs for university students.

Development from Education for Knowledge and Skills to Education for Competencies

The changed ideas concerning instruction, learning, and assessment for competencies resulted in many educational innovations aimed at competency development. In the Netherlands, these changes are seen in all types and levels of education, but vocational education is the pioneer in this respect. Competency-based education is the leading paradigm for innovations in vocational education and vocational colleges are changing – are even obliged to change – their curricula from subject matter curricula to integrated curricula focused on acquiring skills relevant for employability and life-long learning (Biemans et al., 2004; Onstenk, 1997; Tillema, Kessels, & Meijer, 2001).

Development from a Testing Culture to an Assessment Culture: Authentic Assessments

Authentic assessments are gaining field in different types of education (Cummings & Maxwell, 1999; Henderson & Karr-Kidwell, 1998). However, since authentic assessments aim at better preparing students for the labour market, these kinds of assessment seem of most concern for final (i.e., for a job) types of education such as vocation education. Where academic education mainly focuses on theoretical development, and primary and secondary schools on preparing students for further education, vocational education or professional development programmes are inherently more work-oriented and focused on preparing students for the workplace of today (Velde, 1999). As a result, practical demonstration of knowledge and competencies are emphasised as assessment measures in vocational education and authentic assessments mesh well with these ideas (Kerka, 1995). In addition, students in vocational education or professional development programmes are often more practically/work-oriented instead of theoretically oriented. Authentic assessments that show a clear link with professional practice are argued to be crucial for influencing and motivating the learning of these more practically-oriented learners (Huang, 2002; Kasworm & Marienau, 1997).

The Role of Beliefs and Previous Experiences

In vocational education in the Netherlands, assessment practices are developing towards more collaborative activities in which also students and the work field are involved. Much emphasis is placed on involving organisations and corporate enterprises in assessment practices. They are involved in the development of new qualification structures for vocational education based on the jobs in these organisations (Tillema et al., 2000). Furthermore, practitioners are involved in the implementation and actual use of authentic assessments. Work-based assessments, or assessments in the workplace, are gaining popularity and schools more often consult with practitioners to come to a judgement concerning the job performance of a student.

Additionally, increasing emphasis is being placed on integrating learning and working by allowing students to begin their internships as early as possible in their educational trajectory (Ministry of Education, Culture and Science, 2005), especially in practically-oriented, vocational types of education. Here, learning and working in professional practice are alternated on a more regular basis in order to assure a better transition from school to work after finishing school (Tillema et al., 2000). This means that students gain a lot of practical experience during their educational careers.

In short, in VET in the Netherlands, people with different previous experiences are involved in developing and using authentic assessments.

Overview of this Thesis

Between the lines of the general introduction in this chapter, a close reader could have inferred a model that portrays several important variables involved in authentic assessment and the expected influences of these variables on each other (see Figure 1.1). Based on this hypothesised model the studies in this thesis can be described.

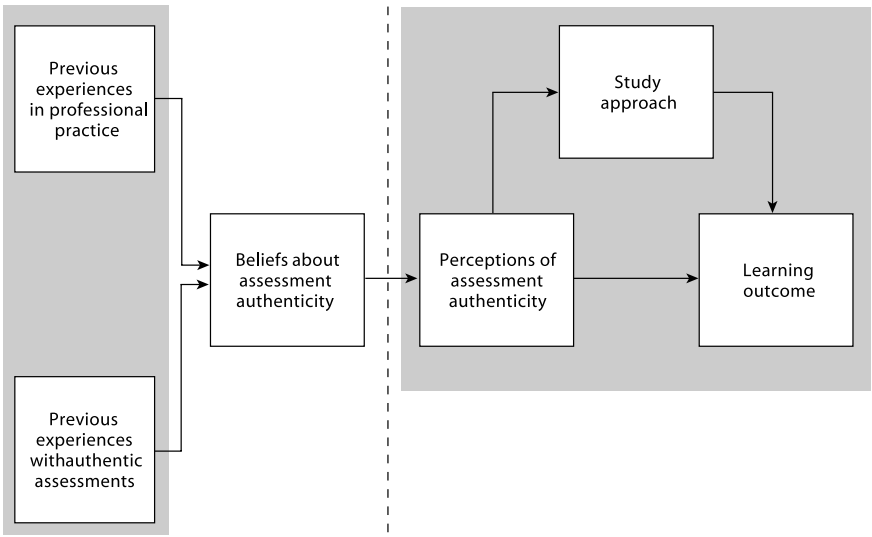


Figure 1.1. Relevant variables for authentic assessment and their relationships

Chapter 2 addresses the question: "What is authenticity?" from an objective and theoretical perspective. It describes a literature study on authenticity and authentic assessment and tries to come to a clear and workable definition of assessment authenticity. This review also resulted in a five-dimensional framework (5DF) for describing assessment authenticity. According to this model, assessment authenticity depends on five facets of an assessment, further subdivided into several characterising elements. The five facets are (1) the assessment task, (2) the physical context, (3) the social context, (4) the assessment form, and (5) the assessment criteria. The 5DF forms the basis for the rest of the studies that examine the beliefs and perceptions about authenticity of various stakeholder groups or the influences of student perceptions on their study approach and learning outcomes.

The studies in chapters 2 and 3 examined whether the hypothesised facets of the 5DF were recognised in practice. In chapter 2 this was done on the basis of an explorative, mostly qualitative study, which examined student and teacher beliefs and perceptions of authentic assessment. The goals of this explorative study were (1) to find out whether the framework covered all important facets of authenticity, (2) to find indications on the relative importance of the five dimensions and (3) to find out if the participant groups differed in their beliefs or perceptions of the facets of authenticity. Nursing student groups with little or much experience in professional practice and with or without experience with authentic assessments in the workplace participated in this study, along with their teachers. Chapter 3 adopts a quantitative approach to determine whether students and teachers perceive and value the facets of authenticity as described in the 5DF. For this purpose, a questionnaire was developed a questionnaire based on the 5DF covering the five dimensions and their characterising elements

of the framework. Results from both studies are used to refine both the 5DF and the perception questionnaire for future studies.

Chapter 4 examines the beliefs about authentic assessment, as a multidimensional construct, from the perspectives of three stakeholder groups involved in authentic assessment practices, namely students, teachers, and practitioners. The students were freshman nursing students who were used to authentic assessment in the workplace. Differences and similarities between these three groups were brought to light through focus group discussions, individual interviews and a questionnaire.

Chapter 5 examines whether students and teachers differ in their perceptions of the authenticity of the five dimensions described in the 5DF when they are confronted with the same kind of authentic assessment. Subsequently, it investigates whether freshman and senior students who differ in their amount of practical experience and authentic assessment experience also differ in their perceptions of the authenticity of the same assessment characteristics. This study was carried out with the adapted questionnaire for measuring perception of authenticity of the five dimensions of the 5DF.

Chapters 6 and 7 both focus on the influence of student perceptions of authenticity and alignment, on their study approach and professional skill development and/or their grades. Chapter 6 zooms in on these relationships by using correlations and structural equation modeling, to examine the direct as well as the indirect effects of the perceptions on deep and surface studying and on generic skill development. This study was conducted *within* one senior student group to get insight into the influence of varying degrees of perceived authenticity on deep learning and generic skill development.

Chapter 7 compares these relationships between perceptions, study approach and generic skill development *between* a freshman and a senior student group to find out if these relationships are stable during an educational career in which students gain a lot of experience in professional practice. Moreover, the authenticity perceptions of both groups are compared as well as their reported study approach and development of generic skills in response to the same kind of authentic assessment. Quantitative data were complemented with qualitative data to gain a deeper insight into the differences or similarities between both student groups in terms of what kinds of operationalisations of the authenticity dimensions student groups perceived as authentic and effective for their learning

Finally, in chapter 8 the results of all studies are topically reviewed in a reflection on the five dimensions of the framework through the eyes of the different beholders participating in the different studies. Differences and similarities between the groups are used to reflect on the hypothesised relationships in Figure 1.1 and to formulate guidelines for developing authentic assessments, based at the 5DF, for student groups with different previous experiences. To conclude, directions for future research are formulated.

Chapter 2

A Five-Dimensional Framework for Authentic Assessment[1]

Authenticity is an important element of new modes of assessment. The problem is that what assessment authenticity really is, is unspecified. This chapter first presents a literature study on authenticity of assessments along with a five-dimensional framework for designing authentic assessments with professional practice as the starting point. This framework was then the subject of a qualitative study to determine whether it was complete and what the relative importance of the five dimensions are in the perceptions of students and teachers of a Vocational Education and Training college for nursing. Implications for the framework are discussed along with important issues that need to be considered when designing authentic assessments.

[1] This chapter is based on Gulikers, J. T. M., Bastiaens, Th. J., & Kirschner, P. A., (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research & Development, 52*, 67-85.

It is widely acknowledged that in order to meet the goals of education, a constructive alignment between instruction, learning and assessment (ILA) is necessary (Biggs, 1996; "constructive alignment theory"). Traditional frontal classroom instruction for learning facts, assessed through short-answer or multiple-choice tests, is an example of just such an alignment. The ILA-practices in this kind of education can be characterised as: instructional-approach: knowledge transmission; learning-approach: rote memorisation; and assessment-procedure: standardised testing (Birenbaum, 2003). This approach to assessment is also known as the testing culture (Birenbaum & Dochy, 1996) and consists primarily of decontextualized, psychometrically designed items in a choice-response format to test for knowledge and low-level cognitive skill acquisition. They are primarily used in a summative way to differentiate between students and rank them according to their achievement. However, the alignment compatible with present day educational goals has changed over the years. Current educational goals focus more on developing competencies relevant for the dynamic world of work than on the acquisition of factual knowledge and basis skills. The ILA-practices that characterise these goals are: instruction that focuses on learning and competence development; learning based on reflective-active knowledge construction; and assessment that is contextualized, interpretative and performance-based (Birenbaum, 2003). This renewed focus of education requires alternative methods of assessments since standardised, multiple-choice tests are widely criticised to not being suitable for assessing higher-order thinking skills or competencies (Birenbaum & Dochy; Glaser & Silver, 1993; Frederiksen, 1984; Segers, Dochy, & Cascallar, 2003). New methods of assessment need not only serve a summative function, they also serve a formative goal of promoting and enhancing student learning. Moreover, they are characterised as being student-centred, meaning that they relinquish more responsibility to the student and increase the involvement of students in the assessment development process. They involve interesting real-life and integrated tasks and authentic contexts as well as multiple assessment moments and methods to reach a profile score for determining student learning or development (Segers, 2003).

The need to contextualize assessment in interesting, real-life and authentic tasks is described as *one of the crucial elements* of alternative, or competency-based assessment suitable for current educational goals (Birenbaum & Dochy, 1996). Increasing the authenticity of an assessment is expected to have a positive influence on student learning and motivation (e.g., Herrington & Herrington, 1998; McDowell, 1995; Sambell, McDowell, & Brown, 1997). Authenticity, however, remains a vaguely described characteristic of assessment, because it is thought to be a familiar and generally known concept that needs no explicit defining (Petraglia, 1998). This chapter focuses on defining authenticity in competency-based assessment, without ignoring the importance of other characteristics of alternative assessment.

Based upon a literature study, a theoretical framework consisting of five dimensions of assessment that can vary in their degree of authenticity is presented. After the description of this framework, the results of a qualitative study are discussed. This study explored student and teacher beliefs about authentic assessment to find out whether the theoretical framework completely describes authenticity or whether important elements are missing. Additionally, the

study explored the relative importance of the dimensions in the perceptions of students and teachers at a nursing college.

The Importance of Authentic Competency-Based Assessment

Authenticity is seen as a crucial quality criterion for valid competency-based assessments for two reasons (Gielen, Dochy, & Dierick, 2003; Messick, 1994). It is expected to be important for the construct validity and for the impact of assessment on student learning or competency development, also called consequential validity. *Construct validity* of an assessment is related to whether an assessment measures what it is supposed to measure. With respect to competency assessment this means that tasks must appropriately reflect the competency that needs to be assessed, that the content of an assessment involves authentic tasks that represent real-life problems of the knowledge domain assessed, and that the thinking processes that experts use to solve the problem in real life are also required by the assessment task (Gielen et al., 2003). Messick (1994) argues that increasing the authenticity of an assessment counters the issue of construct underrepresentation, which is one of the major threats to construct validity. Authenticity, he argues, deals with not leaving anything out of the assessment of a certain construct, leading to minimal construct underrepresentation. In competency-based assessments, authenticity is needed as a quality criterion for assessment to prevent the assessment from becoming an underrepresentation of what happens in the world of work. As a result, authentic competency-based assessments are expected to have higher construct validity for measuring competencies than so-called objective or traditional tests.

Consequential validity describes the intended and unintended effects of assessment on instruction or teaching (Biggs, 1996) and student learning (Dochy & McDowell, 1998). Biggs' (1996) theory of constructive alignment stresses that effective education requires instruction, learning and assessment to be compatible, because these three elements interact with one another instead of function in isolation. If students perceive a mismatch between the messages of the instruction and the assessment, a positive impact on student learning is unlikely (Segers, Dierick & Dochy, 2001). This impact of assessment on instruction and on student learning is corroborated by researchers as Frederiksen (1984; "the real test bias"), Prodromou (1995; "backwash effect"), Gibbs (1992; "the tail wags the dog"), and Sambell and McDowell (1998; "hidden curriculum"). Real test bias and the backwash effect imply that tests have a strong influence on what is taught, because teachers teach to the test, even though the test might focus on things the teacher does not find most important. The tail wags the dog phenomenon emphasises that student learning is largely dependent on the assessment and on student perceptions of the assessment requirements. The hidden curriculum holds that the effects of instruction and assessment on learning are largely based on teacher and student perceptions of the curriculum (i.e., instruction and assessment), which can deviate from the actual intentions of the curriculum. All four ideas support the proposition that assessment strongly influences both learning and instruction. To change student learning in the direction of competency development, authentic, competency-based instruction aligned to authentic, competency-based assessment is needed.

Authentic assessment is expected to positively influence student learning in two ways (Gielen et al., 2003; Herrington & Herrington, 1998; Newmann, 1997). It is expected to stimulate a deep study approach aiming at understanding and application and to stimulate the development of professional competencies. Moreover, it is likely to increase student motivation to learn through the fact that authentic assessments show a direct link to working or social life, outside of school. As a result, students are expected to experience authentic assessments as more interesting and meaningful, because they realise the relevancy and usefulness of it for their future lives (Dochy & Moerkerke, 1997; "task-value of assessment").

Defining Assessment Authenticity

The question that remains is, what is authenticity? Different researchers have different opinions about authenticity. Some see authentic assessment as a synonym to performance assessment (Hart, 1994; Torrance, 1995), while others argue that authentic assessment puts a special emphasis on the realistic value of the task and the context (Herrington & Herrington, 1998). Reeves and Okey (1996) point out that the crucial difference between performance assessment and authentic assessment is the degree of *fidelity* of the task and the conditions under which the performance would normally occur. Authentic assessment focuses on high fidelity, whereas this is not as important an issue in performance assessment. These distinctions between performance and authentic assessment indicate that every authentic assessment is a performance assessment, but not vice versa (Meyer, 1992).

Messick (1994) focuses our attention to the fundamental ambiguity that pervades all authentic assessment practices, namely, *authentic to what*? Honebein, Duffy and Fishman (1993) strengthen the importance of this question by saying that authenticity is a relative concept. In other words, the authenticity of something can only be defined in relation to something else. A test can either be authentic to the school or to the real world. For example, an assessment task can be authentic with respect to school problems, but inauthentic with respect to everyday life experience, because school problems do not relate to everyday life. The point taken in this study is that the authenticity of an assessment is defined by its resemblance to the real world, specifically, to the professional real world. Because current educational goals stress the importance of developing competent employees, we argue that it is important to design assessments that resemble situations that starting professionals or trainees can be confronted with in working life. The situation, according to which the authenticity of an assessment in this study is defined, is called *criterion situation*. A criterion situation reflects a real-life situation that students can be confronted with in their internship or future professional life, which serves as a basis for designing an authentic assessment.

Another issue in defining authentic assessments that logically follows from the previous section deals with *what* students need to learn or develop from working with authentic assessments that resemble professional, real-life situations. Savery and Duffy (1995) define authenticity of an assessment as the *similarity* between the cognitive demands - the thinking required – of the assessment and the cognitive demands in the criterion situation on which the assessment is based. In other words, students need to develop professional thinking skills.

Darling-Hammond and Snyder (2000) argue that dealing only with the thinking required is too narrow, because real-life demands the ability to integrate and coordinate knowledge, skills, and attitudes, and the capacity to apply them in new situations. In their view, authentic assessment includes opportunities for the development and examining of thinking *and* actions. This implies that authentic assessment requires students to *demonstrate* their learning. Birenbaum (1996) deepens this idea of assessing thinking *and* action by emphasising that students not only need to develop cognitive competencies such as problem solving and critical thinking, but also meta-cognitive competencies such as reflection and social competencies such as communication and collaboration. In other words, real life (reflected in the criterion situation) involves different kinds of competencies that all should be taken into account in designing authentic assessments for developing competent employees.

Another crucial point in defining authenticity is the operationalisation of authenticity as a continuum (Newmann & Wehlage, 1993). It is a misconception that something is either completely authentic or completely not authentic (Cronin, 1993). An assessment can be more or less authentic by resembling professional practice to a more or lesser extend.

The definition of authentic assessment used in this study is: *an assessment that requires students to use the same competencies, or combinations of knowledge, skills and attitudes that they need to apply in the criterion situation in professional life*. The level of authenticity of an assessment is defined by *its degree of resemblance to the criterion situation*. This idea is extended and specified by the theoretical framework that describes that an assessment can *resemble a criterion situation along a number of dimensions*.

A complicating matter here is the fact that authenticity is subjective (Honebein, Duffy & Fishman, 1993; Huang, 2002; Petraglia, 1998). This means that people can differ in what they believe to constitute authenticity as well as differ in how they perceive the authenticity of an encountered assessment. This implies that what students perceive as being authentic is not necessarily the same as what teachers and assessment developers see as authentic. If authenticity is indeed subjective, then the fact that teachers usually develop authentic assessments according to their own view causes a problem, namely: although we may do our best to develop authentic assessments, this may all be for nothing if the learner does not perceive it as such. This process, known as "pre-authentication" (Petraglia, p. 53), can be interpreted either as that it is impossible to design an authentic assessment, or that it is very important to carefully examine the experiences of the users of the authenticity of the assessments, before designing authentic assessments (Nicaise, Gibney & Crane, 2000). We choose for the latter interpretation.

Figure 2.1 describes a general framework for the place of authentic assessment in educational practices. It shows that:

1. In light of the constructive alignment theory (Biggs, 1996) authentic assessment should be aligned to authentic instruction in order to positively influence student learning.
2. Authenticity is subjective, which makes student beliefs and perceptions important for authentic assessment to influence learning. Students' beliefs and perceptions might mediate the effect of authentic assessment and/or instruction on student learning.

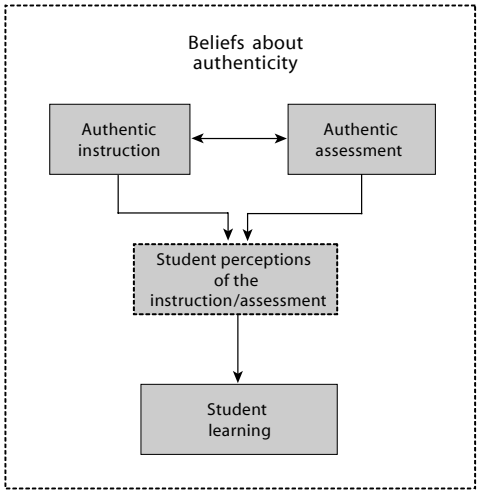


Figure 2.1. A general framework for authentic educational practices

The following section discusses five dimensions (a theoretical framework) that determine the authenticity of an assessment. The purpose of this framework is to shed a light on the concept of assessment authenticity and to provide guidelines for implementing authenticity elements into competency-based assessment.

Towards a Five-Dimensional Framework for Authentic Assessment

As stated, there is confusion and there exist many differences of opinions about what authenticity of assessment really is and which assessment elements are important for authenticity. To try to bring some clarity in this situation, the literature was reviewed to explicate the different ideas about authenticity, authentic assessment and student perceptions of (authentic) assessment elements. Many sub-concepts and synonyms came to light, which were conceptually analysed and divided into categories, resulting in five main facets of authenticity, namely: the task, the physical context, the social context, the assessment result or form, and the criteria. These five facets could be subdivided into several characterising elements.

Moreover, the notion of authenticity as a continuum (Cronin, 1993; Newmann & Wehlage, 1993) resulted in a conceptualisation of the five assessment facets as dimensions that can vary in their degree of authenticity. The degree of authenticity is not solely a characteristic of the assessment chosen; it needs to be defined in relation to the (professional) situation in real-life. For example: carrying out an assessment in a team is authentic *only* if the chosen assessment task is also carried out in a team in real life. Figure 2.2 depicts our five-dimensional framework (5DF) of assessment authenticity. This argues that authenticity is multidimensional and that the degree of authenticity of an assessment depends on the degree of resemblance between these five assessment dimensions and the criterion situation on which the assessment is based. In the next

section, the five dimensions of assessment authenticity and their characterising elements will be thoroughly discussed.

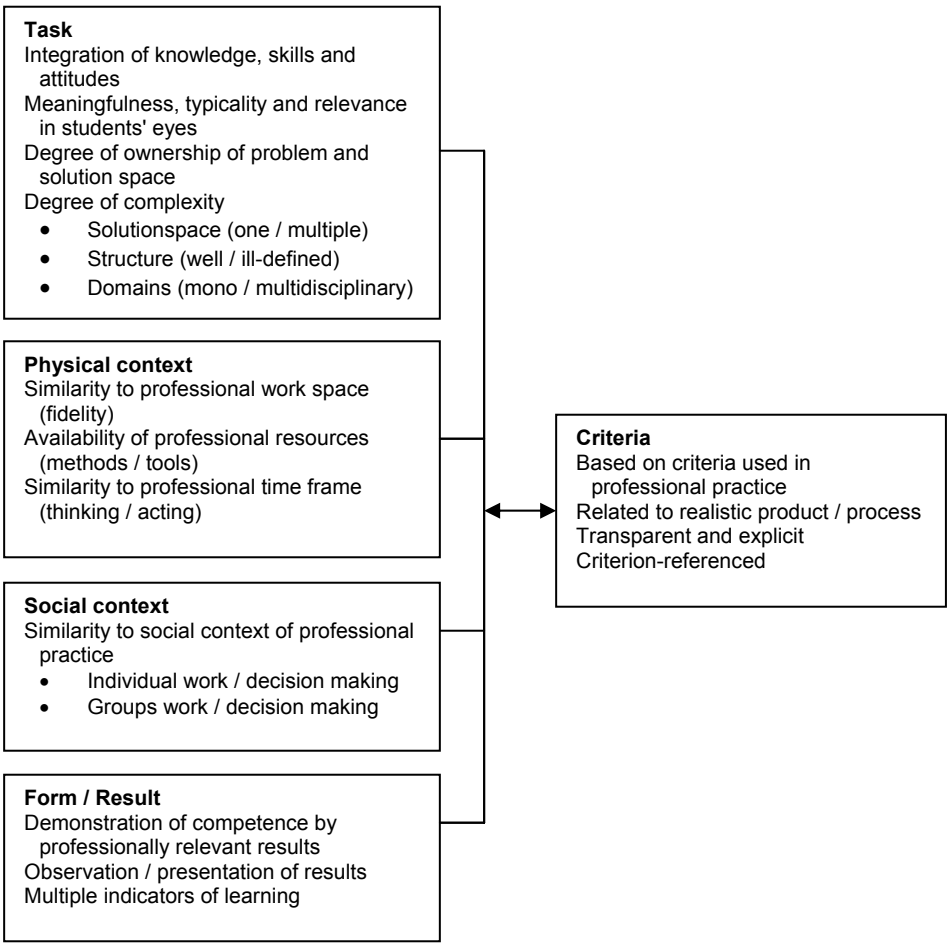


Figure 2.2. The five-dimensional framework for assessment authenticity

Task

An authentic task is a problem task that confronts students with activities that are also carried out in professional practice. The fact that an authentic task is crucial for an authentic assessment is undisputed (Herrington & Herrington, 1998; Newmann, 1997; Wiggins, 1993), but different researchers stress different elements of an authentic task. Our framework defines an authentic task as a task that resembles the criterion task with respect to the integration and use of knowledge, skills and attitudes, its complexity and its ownership. Furthermore, the users of the assessment task should perceive the task, including above elements, as representative, relevant and meaningful.

An authentic assessment task requires students to *integrate and use knowledge, skills and attitudes* as professionals do (Darling-Hammond & Snyder, 2000; Gielen et al., 2003; Van Merriënboer, 1997). Furthermore, the assessment task should resemble the *complexity* of the criterion task (Petraglia, 1998; Uhlenbeck, 2002). This does not mean that every assessment task should be very complex as is often argued by advocates of authentic assessments (e.g., Herrington & Herrington, 1998; Wiggins, 1993). Even though most authentic problems are complex, involving multidisciplinarity, ill-structuredness, and having multiple possible solutions, real-life problems can also be simple, well-structured with one correct answer and requiring only one discipline (Cronin, 1993). The same need for resemblance holds for *ownership* of the task and of the process of developing a solution (Honebein et al., 1993). Ownership for students in the assessment task should resemble the ownership for professionals in the criterion task. Savery and Duffy (1995) argue that giving students ownership of the task and the process to develop a solution is crucial for engaging students in authentic learning and problem solving. On the other hand, in real-life, assignments are often imposed by employers and professionals often use standard tools and procedures to solve a problem, both decreasing the amount of ownership for the employer. Therefore, the theoretical framework argues that in order to make students competent in dealing with professional problems, the assessment task should resemble the complexity and ownership levels of the criterion situation.

Up to this point, task authenticity appears to be a fairly objective dimension. This objectivity is confounded by Messick (1994) and Stein, Isaacs, & Andrews (2004), who argue that student perception of *meaningfulness* of the assessment is at the heart of authenticity. They stress that merely providing assessment tasks representative of the professional discipline is not enough for creating an authentic experience, as long as students do not perceive the assessment as meaningful. Sambell, McDowell and Brown (1997) and Lizzio and Wilson (2004a;b) who showed that it is crucial that students perceive a task as *relevant and representative* of their future professional roles, meaning that (a) they see the link to a situation in the real world or working situation; or (b) they regard it as a valuable transferable skill. Clearly, perceived relevance or meaningfulness will differ from student to student and will possibly even change as students gain experience in professional practice (Lizzio & Wilson, 2004a).

Physical Context

Where we are, often if not always, determines how we do something, and often the real place is dirtier (literally and figuratively) than safe learning environments. Think, for example, of an assessment for auto mechanics for the military. The capability of a soldier to find the problem in a non-functioning jeep can be assessed in a clean garage, with the availability of all the possibly needed equipment, but the future physical environments will possibly involve a war zone, inclement weather conditions, less space, and less equipment. Even though the task itself is authentic, it can be questioned whether assessing students in a clean and safe environment really assesses their capacity to wisely use their competencies in real-life situations.

The physical context of an authentic assessment should reflect the way knowledge, skills and attitudes will be used in professional practice (Brown, Collins & Duguid, 1989; Herrington & Oliver, 2000). *Fidelity* is often used in the context of computer simulations, which describes how

closely a simulation imitates reality (Alessi, 1988). Authentic assessment often deals with high-fidelity contexts. The presentation of material and the amount of detail presented in the context are important aspects of the degree of fidelity. Likewise, an important element of the authenticity of the physical context is that the *number and kinds of resources available,* which mostly contain *relevant as well as irrelevant information* (Herrington & Oliver, 2000), should resemble the resources available in the criterion situation (Arter & Spandel, 1992; Segers, Dochy, & De Corte, 1999). For example, Resnick (1987) argues that most school tests involve memory work, while out-of-school activities are often intimately engaged with tools and resources (calculators, tables, standards), making these school tests less authentic. Segers and colleagues (1999) argue that it would be inauthentic to deprive students from resources as professionals also rely on resources. Another important characteristic crucial for providing an authentic physical context is the *time* students are given to perform the assessment task (Wiggins, 1989). Tests are normally administered in a restricted period of time, for example two hours, which is completely devoted to the test. In real life, professional activities often involve more time scattered over days or on the contrary, require fast and immediate reaction in a split second. Wiggins (1989) says that an authentic assessment should not rely on unrealistic and arbitrary time constraints. In sum, the level of the authenticity of the physical context is defined by the resemblance of these elements to the criterion situation.

Social Context

Not only the physical context, but also the social context influences the authenticity of the assessment. Brown, Collins and Duguid (1989) argued that an authentic activity should reflect practices of the culture or community. In real life, working together is often the rule rather than the exception and Resnick (1987) emphasises that learning and performing out-of-school mostly takes place in a social system. Therefore, a model for authenticity should consider social processes that are present in real-life contexts. What is really important in an authentic assessment is that the social processes of the assessment resemble the social processes in an equivalent situation in reality. At this point, this framework disagrees with literature on authentic assessment that defines collaboration as a characteristic of authenticity (e.g., Herrington & Herrington, 1998). Our framework argues that if the real situation demands *collaboration*, the assessment should also involve collaboration, but if the situation is normally handled *individually*, the assessment should be individual. When the assessment requires collaboration, processes like *social interaction, positive interdependency* and *individual accountability* need to be taken into account (Slavin, 1989). When, on the other hand, the assessment is individual, the social context should stimulate some kind of *competition* between learners.

Assessment Result/Form

The assessment result/form is related to the kind and amount of output of the assessment task, independent of the content of the assessment. In the framework, an authentic result/form is characterised by three elements. It should require students to *demonstrate* their learning or competencies by creating a quality product or performance that they can be asked to produce in real life. In addition, this should be *observable* for others (Wiggins, 1993). Students have to be able

to present to others that their results reflect genuine mastery of the required competencies. The rationale behind requiring students to demonstrate their learning through an observable performance in a real-life situation is that this permits making inferences, as validly as possible, about underlying competencies and predicting future functioning in comparable work situations (Darling-Hammond & Snyder, 2000; Klarus, 2003). Since the demonstration of relevant competencies is often not possible in one single test, an authentic assessment should involve a *full array of tasks and multiple indicators of learning* in order to come to fair conclusions about (professional) competence (Darling-Hammond & Snyder). Uhlenbeck (2002), for example, showed that a combination of different assessment methods was needed to adequately cover the whole range of professional teaching behaviour.

Criteria

Criteria are those characteristics of the assessment result that are valued; standards are the level of performance expected from various grades and ages of students (Arter & Spandel, 1992). Criteria and standards should *concern the development of relevant professional competencies* and should be *based upon criteria used in the real-life (i.e., criterion) situation* (Darling-Hammond & Snyder, 2000). Moreover, some criteria should be *related to a realistic outcome*, explicating characteristics or requirements of the product, process, performance or solutions that students need to create. Setting criteria and making them *explicit and transparent to learners* beforehand is important in authentic assessment (Darling-Hammond, 1994), because this guides learning (Sluijsmans, 2002) and after all, in real life, employees usually know on what criteria their performances will be judged. Moreover, this implies that authentic assessment requires *criterion-referenced judgment*.

Figure 2.2 shows that the criterion dimension has a special status in the 5DF. This dimension has a reciprocal relationship with the other dimensions. On the one hand, criteria based on professional practice, which is often the starting point for developing authentic assessments, should guide the *interpretations* of the other four dimensions. On the other hand, criteria should also reflect the interpretation of another dimension of the framework. For example, if the physical context requires the use of certain resources and tools, the criteria should specify how these should be used in the demonstration of competence, as these criteria guide student learning.

Some Considerations

What does all of this mean when teachers/instructional designers try to develop authentic assessments? What do they need to consider?

The first consideration deals with *predictive validity*. If the educational goal of developing competent employees is pursued, then increasing the authenticity of an assessment will be valuable. More authenticity is likely to increase the predictive validity of the assessment because of the resemblance between the assessment and real professional practice. However, one should not throw the baby out with the bath water (Dochy & McDowell, 1997). Objective tests are very useful for certain purposes as high-stakes summative assessment on an individual achievement, where predicting a student's ability to function competently in future professional practice is not

the purpose. Hence, the purpose of the assessment determines the importance of making an assessment more authentic.

Furthermore, authenticity is only *one* of the (crucial) elements of alternative assessments, next to a number of other important quality criteria for good or valid alternative, performance-based or competency assessment (Dierick & Dochy, 2001; Linn, Baker, & Dunbar, 1991). The framework, as it is described in the theoretical argumentation, describes an ideal picture of authentic assessment practices. In real educational practice, one has to deal with other quality criteria (e.g., reliability) and practical possibilities as well. For example, a certain criterion situation describes that an authentic assessment should allow students to work on the assessment task for several hours spread over one week, while practical possibilities or reliability or accountability considerations make it impossible to completely comply with this timeframe. Every educational assessment requires a compromise between different quality criteria, goals and practical possibilities. However, we argue that increasing the authenticity of the assessment has to be carefully considered in this debate, especially when it comes to competency-based assessments.

Another consideration in designing authentic assessment is that we should not lose sight of the educational level of the learners. Students who are at the beginning of their studies possibly cannot deal with the authenticity of a real, complex professional situation. If they are forced to do this, it will result in cognitive overload and in turn will have a negative impact on learning (Sweller, Van Merriënboer, & Paas, 1998). This implies that the criterion situation, on which the assessment is based, will often need to be a simplification of real professional practice in order to be attainable for students at a certain educational level. The question that immediately comes to mind in this context is "How do you create an authentic assessment for students who are not prepared to function as beginning professionals?". The answer is that the authenticity of an assessment needs to be defined by its degree of resemblance to the criterion situation (i.e., a professional practice situation that the student at his/her educational level can be confronted with in internships), which is not necessarily the same as a professional practice situation at the expert level. Van Merriënboer (1997) argues that a simplification of real professional practice (i.e., the criterion situation) can still be authentic as long as the simplified situation requires students to integrate knowledge, skills, and attitudes, or constituent skills, into a professionally relevant performance. The more students reach the end of their studies, the more the criterion situation will be exactly the same as the real professional practice situation. Thus, a criterion situation is reflection of a professional practice situation at the students' educational level. In this light, criterion situations, on which the assessments are based, are the bridge between learning and working

The final consideration that also sheds a light on the question of what authenticity actually is, is the *subjectivity* of authenticity. The perception of what authenticity is may change as a result of educational level, personal interest, age, or amount of practical experience with professional practice (Honebein, Duffy & Fishman, 1993; Lizzio & Wilson, 2004a;b). This implies that the five dimensions that are argued in the framework for assessment authenticity are not absolute but rather variable. It is possible that assessing professional competence of students in their final

year of study, when they have often done internships and have a better idea of professional practice, requires more authenticity of the physical context than when assessing first year students, who often have little practical experience. Designers must take student perspectives and the changes therein into account when designing authentic assessment.

The exploratory, qualitative study described in the rest of this chapter has two main goals. First, it explores student and teacher beliefs about authentic assessment to find out whether our 5DF completely describes authenticity or whether important elements are missing. Second, it explores the relative importance of the five dimensions. Both these issues are examined in teachers and student groups with different amounts of practical and assessment experience. The differences and similarities between these groups along a limited number of dimensions of authenticity can give insights in what is crucial for defining and designing authentic assessments.

Method

Participants

Students and teachers from a Vocational Education and Training (VET) nursing college took part in this study. One session of the study involved only teachers, one session involved freshman students (second-year) and one session involved senior students (fourth-year). Freshman and senior students differed in the degree of experience in professional practice and freshman were only familiar with (traditional) school assessments, while senior students both experienced school and workplace assessments. The student groups could be further divided into a group of students studying nursing in a Vocational Training Programme (VTP) where they are primarily learning in school and make use of short internships, and a group that studied nursing in a Block Release Programme (BRP) where learning and working are integrated on an almost daily basis. This resulted in five groups of participants: (1) eight freshman VTP students (mean age 18.5 years), (2) eight freshman BRP students (mean age 20.9 years), (3) eight senior VTP students (mean age 19.7 years), (4) four senior BRP students (mean age 31.4 years), and (5) eleven teachers (mean age 42.8 years). The number of participants per session was limited because of the practical possibilities of the Group Support System used in this study.

Materials

An electronic Group Support System (GSS) at the Open University of the Netherlands was used as research tool. A GSS is a computer-based information processing system designed to facilitate group decision-making. It is centred on group productivity through idea generation, preference, and opinion exchange of people involved in a common task in a shared environment. The GSS allows collaborative and individual activities such as brainstorming, idea generation, sorting, rating and clustering via computer communication. To prevent participants (especially students) from feeling inhibited in expressing their ideas and opinions, the GSS was a good option since it is completely anonymous. Furthermore, it was a practical and valuable method because it made it possible to collect a lot of information in a structured way in a short period of time.

To examine the relative importance of the five dimensions, four case descriptions of assessments that varied in their amount of authenticity based on the five dimensions of the model were designed. They described competencies from the nursing competency profile, which were validated by two employees of the nursing college. To check the influence of the GSS-session itself on the perceptions of the authenticity of the cases, the descriptions were used in a pre- and a post-test. To do this, a second set of different but comparable case descriptions were designed, which resulted in two sets of four cases. Cases A and E were completely authentic except for the task; cases B and F were completely authentic except for the physical context; cases C and G were completely authentic except for the result/form; and cases D and H were completely authentic (see Appendix for a full description of a completely authentic case description).

Procedure

All participants had access to a GSS-computer. During a two-hour session, participants carried out both individual and collaborative activities.

At the beginning and end of the GSS-session, participants were presented four case descriptions (ABCD or EFGH). In six paired comparisons (4x3/2), they chose the case that they perceived to be a more authentic assessment. This activity was meant to determine the relative importance of the different dimensions of assessment authenticity in the eyes of the different groups of participants. A second underlying purpose of this activity was to bring participants in a specific reference frame for the rest of the session, and to focus their thinking towards authenticity of assessment instead of assessment in general.

A distinction was made between VTP students and BRP students because it was possible that due to the differences in their studies, they would have different perceptions of what determines authenticity. VTP students, BRP students, and teachers were randomly divided in two "halves"; one which received the cases ABCD in the pre-test and EFGH in the post-test and one which received the cases in the reversed order.

After the initial rating of the case descriptions, the participants were appraised of the purpose of the study. In order to create a specific frame of mind, a very general description was given of the term authenticity (i.e., true to working life). Participants were asked to enter into the system their own statements that described authenticity of an assessment aiming at assessing a student's ability to function in a job. This was a free brainstorm and participants were encouraged to generate as many statements as possible. The statements uncovered what participants believed to be important for assessment authenticity. Statements were anonymously entered into the GSS, where it was also possible to respond to statements made by others. After this electronic brainstorm, the contributions were discussed in order to clarify them. This was recorded for later use and analysis.

Then, a prototype 5DF for authentic assessment was presented as a framework for assessing professional behaviour. The five dimensions were explained to the participants in an attempt to create mutual understanding about the meaning of the dimensions. The five dimensions were

characterised as follows:

1. Task: What do you have to do?
2. Physical context: Where do you have to do it?
3. Social context: With whom do you have to do it?
4. Result/form: What has to come out of it/ What is the result of your efforts?
5. Criteria: How does what you've done have to be evaluated/judged?

After the five dimensions were presented, two additional activities were carried out, which both consisted of paired comparisons to determine the relative importance of the dimensions. The first activity consisted of 10 paired comparisons of the five dimensions (5x4/2). Participants had to choose the dimensions of the framework that they perceived as more important for assessment authenticity. The final activity was the same as the activity at the beginning of the experiment. The participants were again required to carry out paired comparisons of case descriptions that varied in their amount of authenticity according to the 5DF. Each group received the counterbalanced set of case descriptions to those compared at the beginning of the experiment.

Analysis

A characteristic of the GSS is that the answers, statements, or choices of each individual participant are anonymous. This meant that scores per participant were not available. This precluded the possibility of carrying out statistical tests. On the other hand, this anonymity has been shown to stimulate response in idea generation and increase the reliability of answers since socially acceptable answering behaviour is inhibited. The data, thus, were qualitatively analysed. The tapes of the discussions were transcribed. Both discussion statements and the statements keyed in during the brainstorms were analysed as to which of the five dimensions of the framework they fit. Statements that did not fit in one of the five dimensions, were classified as *other*.

The paired comparison data of the five dimensions, that is the number of times that a dimension in the paired comparisons was rated as more important than another dimension, was tallied per participant group. The absolute scores were then translated into rankings. The paired comparisons of the case descriptions were analysed in the same way.

Results

In general, the task, the result/form and the criteria were rated as most important for the authenticity of the assessment. The social context was clearly considered to be least important for authenticity and the importance of the physical context was strongly disputed.

The Relative Importance of the Five Dimensions: Paired Comparisons

The paired comparisons of the dimensions and of the case descriptions gave insight into the relative importance of the five dimensions for designing authentic assessments. The comparisons of the dimensions resulted in five rankings (freshman students VTP and BRP, teachers, senior students VTP and BRP) from 1 to 5. The paired comparisons of the case descriptions were

analysed for the same groups, but were measured in pre- and post-tests, which resulted in ten rankings from 1 to 4.

Table 2.1. Rankings of the five dimensions by the different groups

	N	Task	Physical context	Social context	Result/ form	Criteria
Freshman VTP students	8	2	4.5	4.5	1	3
Freshman BRP students	8	1	3.5	5	3.5	2
Teachers	11	1	4	5	2	3
Senior VTP students	8	2	5	3.5	3.5	1
Senior BRP students	4	2	4	5	1	3
Total	39	8	21	23	11	12

Note. 1 = most important, 5 = least important

Table 2.1 shows rankings per group of the five dimensions based on their perceived importance in providing authenticity to an assessment (1 = most important, 5 = least important). Table 2.1 shows that all groups perceived the task as important (score 1 or 2), while almost all groups, except for the senior VTP-students (score 3.5), perceived the social context as the least important. Furthermore, the result/form and criteria dimensions received more than average importance while all groups perceived the physical context as relatively unimportant (around score 4). In short, independent of the group (see totals in Table 2.1), the task was perceived as most important, followed by the result/form and criteria dimensions; the physical context and especially the social context lagged (far) behind.

The results of the paired comparisons of the case descriptions, in pre- and post-tests, also gave insight into the relative importance of the dimensions. Table 2.2 shows rankings per group of the four case descriptions. A "1" meant that this case was perceived as the most authentic case description and a "4" referred to the least authentic case description. An important finding, for the framework, was that the case that described a completely authentic assessment based on the presence of all five dimensions was perceived as most authentic (score 1) by all, except for the senior BRP students on the post-test (score 2.5). The other three kinds of cases showed an interesting pattern. The case that was authentic except for the task received mostly a score of 2, which meant that this case was perceived as relatively authentic, which in turn meant that the task (which was not authentic in this case) was not perceived as very important in designing an authentic assessment. This is contrary to the findings of the paired comparisons of the dimensions in which the task was perceived as very important in providing authenticity to an assessment. Finally, the participant groups disagreed about the authenticity of the remaining two kinds of cases. All freshman students ranked the case that was authentic except for the result with a "4" meaning that they perceived this case to be the least authentic. In other words, they perceived the result/form dimension as most important for designing an authentic assessment. Teachers, on the other hand, ranked the case that was authentic except for the physical context as the least authentic case (score 4), which meant that teachers perceived the physical context to be

most important in designing an authentic assessment. Senior students did not appear to differentiate, meaning that they perceived the cases with no authentic physical context or with no authentic result/form as equally inauthentic (score 3.5). To sum up, the findings of the paired comparisons of the case descriptions indicated that when all of the dimensions in the framework are present in a case, that the case was seen as the most authentic. Second, there appear to be contradictory results with respect to task authenticity compared to the results of the paired comparisons of the dimensions. Finally, teachers and students appear to differ with respect to the importance of the authenticity of the physical context versus result authenticity when evaluating assessment cases.

Table 2.2. Rankings of the case descriptions by the different groups

	All authentic except for the task	All authentic except for the physical context	All authentic except for the result/form	All authentic
Freshman VTP, pre test	2	3	4	1
Freshman BRP, pre test	2	3	4	1
Freshman VTP, post test	3	2	4	1
Freshman BRP, post test	2	3	4	1
Teachers pre test	3	4	2	1
Teachers post test	2	4	3	1
Senior VTP pre test	2	3.5	3.5	1
Senior BRP pre test	2	3.5	3.5	1
Senior VTP post test	2	3.5	3.5	1
Senior BRP post test	1	4	2.5	2.5

Note. 1 = most authentic, 4 = least authentic

Completeness and Relative Importance: What Do Participants Believe?

Table 2.3 shows that all dimension received attention in the brainstorms and discussions. Furthermore, these results corroborated the earlier findings in that the social context received the least attention in all groups. Besides the five dimensions, almost all their characterising elements of the dimensions, described in the framework, were reviewed.

Based upon the number of statements and the ratios of the statements compared to each other as shown in Table 2.3, freshman students placed primary interest on the task followed by the physical context. Seniors and teachers placed equal emphasis on task and result. Teachers differed from all students, regardless of the year, with respect to the emphasis on the physical context. Teachers devoted a lot of time discussing the required fidelity level of the physical context in an effective authentic assessment. Especially the question whether the physical context should be real professional practice or a simulation in school was discussed. No clear differences were found between VTP and BRP freshman student and VTP en BPR senior students. Therefore, they were treated as one freshman and one senior group.

Table 2.3. Number of statements per dimension of each group

	N	Task	Physical context	Social context	Result/ Form	Criteria	Other
Freshman students	16	24	19	6	7	13	45
Senior students	12	34	21	9	36	12	26
Teachers	11	16	39	5	19	21	56

A closer look at the content of the brainstorm statements gave the impression that teachers and seniors agreed more with each other and with the ideas of the framework, than the freshman students, especially when it comes to the task and the result/form dimensions. Teachers and seniors agreed with the framework that an authentic task required an integration of professional knowledge, skills, and attitudes and they acknowledged that the task should resemble real-life complexity. On the other hand, freshman students were preoccupied with knowledge testing, they had problems picturing the idea of integrated testing, and were primarily concerned with making assessment more clear and easy (e.g., "assignments should be less vague, not more than one answer should be possible") instead of simulating real-world complexity. In the result/form dimension, teachers and seniors agreed that more assessment moments and methods should be combined for a fairer and more authentic picture of a student's professional competencies. Freshman students did not discuss the result/form dimension much; they only mentioned that reshaping current tests in the form of cases would make it more realistic. In other words, every kind of assessment could be made more authentic by adding realistic information.

Table 2.4. Variables in the "other" category per group

	Freshman students (n = 16)	Senior students (n = 12)	Teachers (n = 11)
General statements applicable to all five dimensions	6	1	2
Instruction	28	7	5
Alignment instruction – assessment	2	3	3
Alignment school – practice	6	3	3
Assessor	3	3	6
Organisation/preconditions	-	-	7
Influence on the learning process	-	-	4
Not defined/nonsense	-	9	26

A specification of the *other* statements (see Table 2.4) showed, first, that all groups made statements emphasising the alignment between instruction and assessment and between school and real-life practice. This is in agreement with the theoretical ideas behind the general framework for authentic assessment (Figure 2.1). Second, Table 2.4 showed that issues

concerning the assessor of an authentic assessment and organisational or preconditional issues are important issues in discussions about authentic assessment. Issues related to the assessor dealt with the realisation that people from professional practice should be involved in defining and using criteria and standards. Organisational issues involved statements about conditions that should be met *before* authentic assessment can be implemented in school. For example, teachers talked about placing students in professional practice sooner and more often for the purpose of assessing students in this professional context. Finally, Table 2.4 showed that freshman students took the opportunity to talk and complain about the "instruction". Although this was not asked of them (i.e., it was about assessment) 28 statements dealt with what was taught and not with what was assessed. Seniors were more focused and teachers statements were spread over different *other* variables and the 26 statement of the 'not defined' variable included mostly jokes or questions they asked each other.

Conclusion

Overall, the 5DF gave a good description of what dimensions and elements should be taken into account in an authentic assessment; the participants discussed all dimensions and almost all elements described in the framework. However, organisational or preconditional issues as well as questions dealing with who should be involved in the use and development of the assessment needs to be addressed in the discussion about authentic assessment

A combination of the results of the GSS-activities led to the conclusion that task, result/form and criteria were perceived as very important for assessment authenticity. The physical context was most important in the eyes of the teachers. The social context was perceived as the least important dimension.

Furthermore, not all groups perceived the dimensions and elements in the same way. The teachers and seniors mostly agreed with each other and with the theoretical framework, while the freshman students often deviated from the other groups. There were no differences between VTP and BRP students.

Discussion

At this point it is necessary to restate the perspective of this research. The questions with which we began were: (1) Is the framework complete?; (2) What is the importance of the five dimensions?, and (3) Do students differ from teachers with respect to what they believe or perceive as important for authenticity? These questions shed a light on possible guidelines for designing authentic assessments.

With respect to the first question, the answer appears to be yes. The five dimensions appear to adequately define authenticity as seen in both the brainstorms and the high ranking of those cases that were authentic on all five dimensions. The adequacy of the framework is corroborated by the finding that during the brainstorms most characterising elements of the dimensions, as described by the framework, were seen as important when designing authentic assessment. However, all groups referred to the involvement of professional practice in the development and use of criteria. This is an issue to consider in the criterion dimension of the framework. The

organisational issues addressed in the discussions should not be part of the 5DF as they deal with preconditions that have to be met before the implementation of an authentic assessment can become successful.

Concerning the relative importance of the dimensions, the paired comparisons showed some subtle differences in the importance of the five dimensions for providing authenticity. While the task, the result/form and the criteria dimensions were perceived to be very important for authenticity, the physical context and especially the social context were seen as less important. The social context is unequivocally perceived as the least important dimension of authenticity. All groups stressed the need for individual testing, while on the other hand students as well as teachers stressed that most nursing activities in real life are collaborative. Teachers explained that, "assessing in groups is a soft spot, we just don't know how to assess students together, because at the end we want to be sure that every individual student is competent". It should not be concluded, based on these findings, that the social context is not important for assessment authenticity, but if choices have to be made in designing an authentic assessment, the social context is probably the first dimension to leave out.

The findings on importance of the task are sometimes contradictory. While the brainstorms and the paired comparisons of the dimensions showed that the task was perceived as very important by all, the paired comparisons of the cases made the task seem less important. It is possible, thus, that while the respondents consider the task (as abstracted concept) to be most important, they are not able to identify (i.e., they do not perceive) an authentic task. A possible explanation for this is that the all-authentic-except-for-the-task case resembles current assessment practices. Because previous experiences are found to strongly influence perceptions (Birenbaum, 2003), the familiarity of these cases may have influenced the paired comparisons of the cases. If this was the case, the paired comparisons of the five dimensions were probably a more objective measure of the importance of the five dimensions.

With respect to the third question concerning the *differences between students and teachers* in their beliefs and perception of authenticity, some interesting findings came to light. The most differences were found between the freshman students and the teachers, while the seniors agreed with the teachers more often. Moreover, the beliefs of the teachers and seniors agreed more with the ideas of the theoretical framework. Possibly, the beliefs of the older students have changed during their college career as a result of having had more experience with professional practice and with different kinds of assessments; the beliefs of the freshman students - who have less practical experience and no experience with authentic assessments - seemed to be primarily based on their previous experiences with 'traditional' methods of assessment, which explained the focus on knowledge and in-school testing. In other words, it looks like freshman students have different beliefs and possibly misbeliefs of real professional practice and thus of authenticity.

Furthermore, the brainstorms and the paired comparisons of the case descriptions showed differences between teachers and students in the perception of the physical context. Teachers focused on the importance of increasing the authenticity of physical context by placing the

assessment in professional practice, while students, especially freshman, mostly focused on in-school testing with for example simulation patients and realistic equipment.

Finally, all groups agreed on the relative unimportance of the social context and on the importance of using criteria that resemble the criteria used in real professional practice. Teachers and students agreed that, at this point, the criteria used in school differ too much from criteria used in professional institutes and that school-criteria are often unknown or misinterpreted by assessors at the professional institutes.

Future Implications

The findings of this explorative study allow for some critical questions and guidelines concerning the design of authentic assessment. First, student beliefs and perceptions should be considered in designing effective authentic assessments. The qualitative results of this study showed that students with little practical experience and no experience with assessment at the work floor, had different beliefs (possibly misbeliefs) of what authenticity means than older, more experienced students and teachers. As a result, these groups perceived the importance of the authenticity of the assessment dimensions differently. For authentic assessment to work, two options need to be considered in this matter. Either the assessment meets the beliefs of the freshman students, for example by sticking to explicit knowledge testing in the name of authentic assessment, which is likely to confirm unwanted learning behaviour. Or changing student beliefs and thereby opening the possibilities to change their learning behaviour towards professional competency-development should be given explicit attention when implementing authentic assessment. In addition, the (mis)beliefs of freshman students might indicate that the importance of assessment authenticity, in general, increases as students proceed through their studies.

Second, we might be able to save precious time and money in the design, development and implementation of authentic assessment with respect to the physical context and the creation of social contexts. Previous research (Gulikers, Bastiaens, & Martens, 2005) showed that a more authentic physical context did not automatically lead to improvements in student performance. Research should examine if assessing students in a real professional-context has additional value for students, or if assessing in an (electronic) simulation in school is authentic enough as long as students are confronted with an authentic task, result/form and criteria (see also Gulikers et al., 2005). Simulation in school, virtual or not, is probably easier and less expensive to implement and therefore warrants careful consideration.

The explorative nature of this study without the possibility of quantitative statistical analyses due to the nature of the GSS makes firm conclusions impossible. However, the electronic GSS efficiently delivered a lot of qualitative data in a short period of time. What the data of this study *do* show is that authenticity is definitely a multi-faceted concept and that a number of the facets (dimensions) appear to be of more importance than others. This can have far reaching implications for educational design.

Since authentic assessment should be aligned to authentic instruction, as argued in the beginning of this chapter (Figure 2.1), the five dimensions of the framework might also be applicable to authentic instruction. The 5DF can be used as design model to foster alignment

between instruction, learning and assessment. Learning tasks stimulate and support students to develop the competencies that professionals have and an assessment task asks students to demonstrate these same competencies without additional support (Van Merriënboer, 1997). Schnitzer (1993) stresses that for authentic assessment to be effective, students need the opportunity to practice with the form of assessment before it is used as an assessment. This implies that the learning task must resemble the assessment task, only with different underlying goals. Learning tasks are for learning and assessment task are for evaluating the level of learning in order to improve (formative) or make decisions (summative). The 5DF can deal with a (conceptual) alignment between authentic instruction and assessment.

The actual effectiveness of the framework for designing authentic assessments, however, should be examined by evaluating the influences of different kinds and levels of authenticity of assessment on student learning and motivation. Because implementing authenticity elements in assessment requires a lot of time, money and energy (Martens, Bastiaens, & Gulikers, 2002), research should examine which elements of the framework are crucial for affecting student learning in the direction of the development of professional competencies.

The argumentation of the theoretical framework and the qualitative study gave some interesting impulses to further theoretical and practical research concerning assessments authenticity. All participants in this study agreed that instruction and assessment in school should be aligned with each other and that developing education that focuses on the development of competencies and takes professional practice as a starting point, requires assessments that are also competency-based and based on professional practice. In other words, it requires authentic assessment.

Appendix

Authentic Case Description: Authentic on Five Dimensions

You're working as an intern in a nursing home. You have to show that you are capable of providing basic care to geriatric patients with different kinds of problems. You have to be able to help them with their own personal care such as washing themselves, getting dressed and combing their hair. In school you've learned and practiced what basic care for geriatric patient means, how you should provide this to, and what you should take into account while doing this (e.g., disabilities or dysfunctions, privacy issues and rules). To judge whether you are competent, your mentor at the nursing home observes you while you are taking care of three different patients. The first patient is suffering from dementia, but is physically in good health, the second is paraplegic and the third is a very overweight gentleman. You are allowed to take care of these patients on your own, but you can also ask colleagues for assistance if you think the care is too trying or difficult to carry out alone. The assessment criteria tell you that you are required be sensitive to the different disabilities of the patients and that you take this into account in your actions. Furthermore, you need to communicate well with the patients, explain to them what is going to happen, and answer their questions. Your mentor will write a report on your performance in each of the three cases and finally draw a conclusion about your competence in providing basic care to geriatric patients.

Authentic Task
Provide personal care to geriatric patients (washing, getting dressed, combing) while taking into account their disabilities, dysfunctions and privacy rules and issues.

Authentic Physical Context
Performing the tasks during an internship at a nursing home for geriatric patients.

Authentic Social Context
Allowing the student to take care of the patients on his/her own, but also allowing him/her to also ask for assistance if the task is to trying/difficult to carry out alone.

Authentic Result/Form
Judging the competence of the student by observing the student while he/she is demonstrating the competence through performing the whole task. Competence is judged based on multiple (three) performances in different contexts (different patients).

Authentic Criteria
Judging an integrative whole of nursing competencies by taking different aspects of the competent job performance into account (washing, dressing, combing, but also communication, taking disabilities into account, taking privacy rules into account). Criteria are known to the students – since they learned them in school before beginning the internship – and are based on the basic care that nurses have to provide in real-life.

Chapter 3

The Practical Value of the Five-Dimensional Framework for Assessment Authenticity: Student and Teacher Perceptions[2]

This chapter builds on the five-dimensional framework (5DF). The study reported on examined whether or not the theoretical dimensions are supported in practice by exploring the perceptions of both Vocational Education and Training students and teachers. The framework led to the development of a questionnaire for examining if the dimensions and several characterising elements of the 5DF were recognised by students and teachers in practice. Reliability and factor analysis as well as readability scores were used. Teachers recognised both the dimensions and characterising elements as facets that determine assessment authenticity. In the eyes of the students, four of the five dimensions (Task, Physical Context, Form and Result/Criteria) determine authenticity, while students did not perceive the Social Context as a characteristic of assessment authenticity, neither did they differentiate the characterising elements. Implications for using the 5DF to develop or evaluate authentic assessments are discussed.

[2] This chapter is based on Gulikers, J. T. M., Bastiaens, Th. J., & Kirschner, P. A. (2006). Authentic assessment, student and teacher perceptions: the practical value of the five-dimensional framework. *Journal of Vocational Education and Training, 58*, 337-357.

Assessment, in the past often referred to as "testing", has been an important aspect of educational practice for a long time (e.g., Bloom, 1956) and the idea that assessment is a salient variable in determining what and how students learn has become an often examined subject of research in the last two decades (e.g., Scouller, 1997; Scouller & Prosser, 1994; Thomas & Bain, 1984). Segers, Dochy, and Cascallar (2003) argue that several aspects characterise tests or assessments, one of which being the authenticity continuum. This continuum shows that an assessment can span the gap between artificial and decontextualized on the one hand or authentic and situated on the other. New modes of assessment, that focus on competencies needed for future jobs, tend to lean towards the authentic side of the continuum, since authenticity is expected to be crucial for preparing students for the dynamic world of work that characterises current society (Boud, 1995; Segers et al., 2003). Moreover, making an assessment authentic is expected to have a positive influence on student learning and motivation (Herrington & Herrington, 1998). Increasing the authenticity of an assessment is expected to stimulate students to develop skills or competencies relevant for their future world of work. But authenticity is not an 'objective' quality as such. Something is only authentic with respect to something else, for example a person, place or thing (Honebein, Duffy, & Fishman, 1993). This means that what one person perceives as authentic is not necessarily authentic in the eyes of someone else. In reality, thus, "authenticity is in the eye of the beholder".

We argue that it is important to explore the concept of authenticity from two different angles, namely a theoretical (objective) and a practical (subjective) angle. The literature study (Gulikers, Bastiaens, & Kirschner, 2004) described in chapter 2 examined assessment authenticity from a *theoretical angle*, which resulted in the five-dimensional framework (5DF) for describing assessment authenticity from an objective viewpoint. The study in this chapter focused on the *practical angle* by examining what determines authenticity in the perception of different users (e.g., developers or assessees). More specifically, it examines if students and teacher recognise and support the theoretical dimensions of authenticity, as described in the 5DF. Boud (1995) argued that assessment research should focus on carefully examining assessments as students see it, since this provides us with the most relevant information for developing assessments that are helpful for student learning.

Theoretical Background

Boud (1990) argued that a major problem of education is the fact that there are gaps between teaching and professional practice and between assessment tasks and what occurs in the world of work. In the last decade of the previous century, the educational culture changed from knowledge-based towards competency-based education and the educational goal became to develop competent students and future employees (Segers et al., 2003). This made the issue of bridging the gaps between learning and working even more salient.

In Vocational Education and Training (VET) in the Netherlands, the context of this study, the link between learning and working is even more crucial, because of several characteristics of this type of education (Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004): (a) VET is primarily a final (for a job) type of education, (b) there is a strong focus on becoming a

practitioner and less on pure theoretical development, and (c) students at these schools are often more practically oriented and directed towards working instead of studying. As a result, authentic learning has become an important issue in these schools (Kerka, 1995). Ideas from cognitive apprenticeship (Collins, Brown, & Newmann, 1989) and situated learning (Brown, Collins, & Duguid, 1989), in which authentic learning plays a central role, have resulted in educational practices in which learning activities are contextualized in realistic situations. Newmann and Wehlage (1993) argued that schools, at least work-oriented schools such as VET colleges, need to provide students with authentic real-life learning experiences, with their complexity and limitations, to stimulate students to more higher-order thinking processes and active learning. However, changing only the instruction is not enough to foster authentic learning. In line with the constructive alignment theory (Biggs, 1996), instruction and assessment *both* need to focus on authentic learning. In other words, it is imperative that assessment practices also change towards more authentic assessment forms.

The need to contextualize assessment in interesting, real-life and authentic tasks is considered *one of the crucial elements* of new modes of assessment (Birenbaum & Dochy, 1996). Dochy (2001) described the assessment of the application of knowledge to actual, real-life (authentic) cases as the core goal of alternative assessments. Gielen, Dochy, and Dierick (2003) even argued that authenticity of the assessment tasks is imperative to achieve the expert level of problem solving. In short, authenticity is argued to be important for preparing students for the unexpected world of work.

Authenticity, thus, seems a crucial characteristic of assessments. The problem, however, is that what is meant by authenticity is often not clearly defined. The literature study described in chapter 2 showed that authenticity can be, and often is, described in very different ways. In addition, no clear guidelines exist for developing authentic assessments. This results in authentic assessment practices that "dress up" existing assessments with some real- world elements (Cummings & Maxwell, 1999), without a clear understanding of what these real-world elements are and how these are appropriately implemented.

The Theoretical Angle: Objective Authenticity

A crucial question in defining authenticity is: "authentic to what?" (Messick, 1994, p. 18). Honebein and colleagues (1993) argued that the authenticity of something can only be defined by its resemblance to *something else* and it is the specification of this something else that is crucial for a further discussion about and examination of the concept of authenticity. Since the goal of education is to prepare students for the world of work, at least in vocational education (Biemans et al., 2004), the point taken in this study is that the authenticity of an assessment should be defined by its *resemblance to students' current or future professional practice*. Resembling professional practice, however, means more than merely implementing some superficial realistic elements, which only leads to, as Cumming and Maxwell (1999) call it, "camouflage" (p. 188). Developing an authentic assessment should start with an analysis of the professional practice situation to find out what kind of knowledge, skills and attitudes (or competencies) experts use when handling this situation. As a result, the authentic assessment should *require students to use*

and demonstrate the same (kind of) competencies, or combinations of knowledge, skills and attitudes, that are applied in this situation in professional life (Gulikers et al., 2004).

Additionally, Cronin (1993) and Newman and Wehlage (1993) argued that authenticity is a continuum. This means that an assessment can be more or less authentic depending on its *degree of resemblance to the professional practice situation*. The theoretical 5DF expands this idea by arguing that an assessment can *resemble professional practice along five dimensions.* This framework describes five assessment characteristics that determine the degree of authenticity, at least from a theoretical viewpoint. It argues that the authenticity of an assessment depends on the resemblance to the professional practice situation of (a) the assessment task(s), (b) the physical context in which the assessment takes place, (c) the social context of the assessment, (d) the result or form that defines the output of the assessment, and (e) the assessment criteria. From now on, these will be referred to as *the five dimensions of authenticity*. These five dimensions can be further described by several characterising elements (see chapter 2 and Figure 2.2 for a full description of this framework).

The main idea behind this framework is that different kinds of authentic assessments can be developed by varying the degree of resemblance between the five dimensions with their characterising elements and the professional situation that the assessment aims to reflect.

Up to this point, one could argue that authenticity seems to be an objective concept. We argue, however, that authenticity is not purely objective, but that an important part of authenticity is in the eye of the beholder, as will be described in the next section.

The Practical Angle: Subjective Authenticity

Just as the concept "expensive" has an *objective side*, as in a Rolls Royce is expensive compared to an Austin Mini-Cooper (the original), it also has a *subjective side* in that a Rolls Royce is expensive for an average person, but not for a billionaire. The same is true for assessment authenticity. It has an objective side, which was just handled, but also a subjective one, namely how the assessee/assessor perceives the authenticity of an assessment.

It is important to investigate this subjective side of authenticity for two reasons. First, student perceptions of assessment characteristics are found to determine what and how students learn (Entwistle, 1991; McDowell, 1995; Scouller, 1997; Struyven, Dochy, & Janssens, 2003; Van Rossum & Schenk, 1984). This implies that not the objective authenticity, but rather student perceptions thereof influence their learning. Before an authentic assessment will stimulate students to develop professionally relevant skills or competencies, students have to perceive the assessment as being authentic with respect to their future professional life. Second, student and teacher perceptions of assessment characteristics are found to differ (MacLellan, 2001; Ngar-Fun, 2005). With respect to authenticity, Honebein and colleagues (1993) argued that its perception can change as a result of age, kind of and amount of schooling or practical experience. As a result, students and teachers are likely to differ in how they perceive authenticity. When this is indeed so, problems for educational practices might arise, since teachers are mostly the ones to develop the authentic assessment and they do so according to what they think is authentic. This process is called "pre-authentication" (Petraglia, 1998, p. 53)and shows the relevance of exploring teacher

as well as student perceptions of authenticity, as it is important to develop assessments that are perceived as being authentic not only by teachers, but by students as well.

This Study

The goal of this study is to examine if the authenticity dimensions of the 5DF are corroborated in practice. It examines if the theoretical 5DF adequately differentiates the determining dimensions of assessment authenticity in the eyes of students and teachers. For this purpose, we needed to develop a way to measure if and how students and teachers value and differentiate various dimensions of authenticity in an assessment. This led to the development of a perception questionnaire based on the 5DF. If different dimensions are perceived, the idea that an assessment can be made more or less authentic in various ways by manipulating those dimensions in an assessment, is supported. On the other hand, if students and/or teachers do not recognise various dimensions, the use of the 5DF in theory and in practice, has to be reconsidered. Hence, this study might provide practical guidelines for developing and evaluating authentic assessments.

Method

Participants

One hundred and fifteen students (mean age = 18.1, *SD* = 3.27) of a Vocational Education and Training (VET) for Social Work and 18 of their teachers enrolled in this study. The students were in their second year of study and studied Social Work in a vocational training programme in which learning and working were alternated on a regular basis. Moreover, students were familiar with the authentic assessment used in this study. The 18 teachers were all assessors in the authentic assessment that was the object of this study.

Materials

The authentic assessment. This study was designed around an authentic assessment on the topic of "determining care needs". Students received a case description of a handicapped client, living in a social home, who has been in a car accident and had physical and behavioural problems afterwards. This caused problems for the client himself as well as for the people around him. The goal was to draw up an activity and guidance plan to help the client and to improve his functioning in the community. The assessment consisted of two parts. First, students received the case description of the client and had to formulate, in writing, their ideas and the actions they were planning to analyse and observe the client. Second, students were confronted with the client in a ten minute role-play in which they had to observe the client and discuss their activity and guidance plans with this client. Both activities took place in school. The second part was a simulated role-play in which a teacher played the role of the client. Every student performed the assessment individually. Moreover, student performance was observed and graded by two assessors (teachers) on a list of ten performance criteria known to the students, and the goal of the assessment was summative.

The instructional phase. A competency-based instructional period of 9 weeks preceded the authentic assessment. This period focused on planning activities within the social work institution. During 8 weeks, students worked in groups on critical professional problem situations, for example "analysing client needs", "planning individual activities based on client needs", "or planning group activities". They had to set learning goals focusing on knowledge as well as skills/attitudes. After a period of self-study and skill-training, all group members had to perform a formative assessment. This was a role-play assignment based on a new, but related problem case. The summative assessment (in this case determining care needs) was based on a selection of course objectives and performance criteria. The performance criteria for the summative assessment were made known to students one week prior to the assessment in which students were freed from obligatory educational activities and in which they could choose for themselves how to prepare for the assessment

The perception questionnaire. To study student and teacher perceptions of authenticity, a 5-point Likert scale perception questionnaire was developed based on the authenticity dimensions of the 5DF. The idea behind the questionnaire was to first measure perceptions at a dimensional level (the resemblance between the five dimensions [task, physical context, social context, form/result, and criteria] and professional practice) and then, to measure the perception of four of their, more specific, characterising elements (the resemblance between the characterising elements and professional practice). In addition to measuring perceptions of the authenticity of different dimensions of the assessment, the Vocational Relevance scale of the Course Perception Questionnaire (Entwislte & Ramsden, 1983) was used to measure perceptions of the overall authenticity of the assessment. This scale implicitly assumes that the authenticity of an assessment is one-dimensional (this assessment [as a whole] is based on professional practice) instead of multidimensional as is presumed by the 5DF. An additional scale was developed to measure the kind of learning that was stimulated by this assessment. This scale was developed to examine if it was indeed perceived that the assessment assessed the capability to apply knowledge and skills to real-life situations.

Table 3.1 shows the scales of the student questionnaire, accompanied by an example of an item and the number of items per scale. The teacher questionnaire was almost identical, except that the word "I" was replaced with the word "the student". All items were contextualized in a specific assessment (here: determining care needs) and referred to students' future professional practice as a social worker (an example of an item: "This assessment task resembled the tasks of a real social worker").

After the items were constructed, four teachers of different VET schools were asked to review the items. They were asked if the items were readable, understandable and clear for students at this educational level. Where necessary, the questionnaire was adapted in line with their suggestions.

Procedure

During one week, all students took part in the assessment. Every student completed the perception questionnaire after finishing the assessment. Teachers filled in the questionnaire at the end of the assessment week.

Table 3.1. The scales of the perception questionnaire

Main scale	Items	Example
1. Task	5	The task of this assessment is an important aspect of a social workers' job
a. Task complexity	4	The task of this assessment was more complicated than the tasks I have to perform in my work placement
b. Task ownership	5	The responsibility that I got in solving this assessment task is different from the responsibility I get in my work placement
2. Physical context	5	The context in which I had to perform this assessment resembles the professional practice of a social worker
c. Resources availability	5	In this assessment I could use of all the resources/ equipment that are used in professional practice
3. Social context	4	If I had to perform this task in my work placement, I would have cooperated more with my colleagues
4. Result/form	6 (3/3)	The result that I had to produce in this assessment was something a social worker has to produce practice also
5. Criteria	4	The criteria that were used in this assessment are different than the criteria that are used in professional practice
d. Transparent criteria	4	It was hard to find out what was expected of me in this assessment
Overall authenticity	5	This assessment was oriented to my future profession of social worker
Kind of learning	5	In this assessment I had to apply the thing I'd learned in a professional practice situation

Analysis

Because the questionnaire was developed based on theoretical insights, confirmatory reliability analyses were done to find out if the scales of the perception questionnaire reliably measure the authenticity facets that were intended. Cronbach Alpha was calculated for all scales, for students and for teachers separately, with $\alpha = .6$ as lower limit. The decision to examine reliabilities for teachers and students separately was guided by the possibility that the reliabilities differ between these groups since the groups represent two populations that differ on various aspects (e.g., age and amount of practical experience and schooling [see Honebein et al., 1993]) that can influence their way of answering questions about authentic assessment.

In addition, an exploratory factor analysis (Principal Component Analysis with Varimax rotation) was conducted on the 115 student questionnaires. The factor analysis was used to:

1. Explore the assumption of multidimensionality of the authenticity concept
2. Explore how students structured the concept of assessment authenticity, and
3. Explore if this factor structure corroborated the authenticity facets of the perception questionnaire and the 5DF.

For further validation, three reviewers were asked to interpret the factors. These reviewers were researchers who were not involved in this study and were both unfamiliar with the scales

of the perception questionnaire and the 5DF. The factors and their interpretations were compared with the theoretical ideas of the 5DF.

Results

Table 3.2 shows the results of the reliability analysis on the scales of the perception questionnaire for students and for teachers.

Table 3.2. The reliabilities of the scales of the perception questionnaire.

	Students (n = 115)	Teachers (n = 18)
Task	.62	.86
Task complexity	.49	.81
Task ownership	.41	.59
Physical context	.80	.89
Resource availability	.51	.83
Social context	.35	.84
Result/form	.72	.71
Criteria	.55	.74
Criteria transparency	.73	.72
Overall authenticity	.70	.96
Kind of learning	.61	.86

This table presents some interesting findings. First, all reliabilities were higher in the teacher group than in the student group. Second, all scales (except for the task ownership sub-scale) were reliable in the teachers group, while 5 out of the 11 scales did not exceed .6 for the student group. Third, at the dimensional level, the task, the physical context and the result/form dimensions were reliable in both groups, while the criterion dimension showed marginal reliability and the social context showed very low internal consistency in the student group. Fourth, at the more specific level of the characterising elements within the dimensions, only the criterion transparency scale was reliable in the student group.

Two explanations could be given for the unreliabilities in the student group: (a) There was no straightforward fit between the scales and the underlying constructs in this questionnaire, resulting in a different clustering of the items than in the pre-defined scales; or (b) the items were too difficult for VET students to understand. Both these options were examined.

First, an explorative factor analysis was done to examine if the underlying factor structure in the questionnaire could explain the unreliabilities. An initial factor analysis resulted in 15 factors possessing eigenvalues of 1.0 or more. However, these factors were impossible to interpret and the scree-plot suggested a six-factor solution as a more appropriate structuring of the student perceptions (Cattell, 1966; Lizzio & Wilson, 2004). When a reliability analysis was conducted per factor, the first six factors turned out the have a Cronbachs alpha of more than .6.

Table 3.3 Factor loading of the final items

	Item	Factors					
		1	2	3	4	5	6
O	This assessment was oriented to my future profession of social work			.64			
O	This assessment was clearly directed to professional requirements	.58					
O	This assessment prepared me for my future professional			.56			
T	The task of the assessment resembled the task of a real social worker			.77			
T	The task of this assessment was an important part of the social work profession			.60			
T	The task of this assessment differed from the tasks of a real social worker			.60			
PC	The context of in which I had to perform the assessment was fake		.79				
PC	The context of in which I had to perform the assessment looked like a social workplace		.66				
PC	The context of in which I had to perform the assessment looked just like the real world		.85				
PC	The context of in which I had to perform the assessment was realistic		.79				
F/R	This way of assessing is an effective way of assessing professional skills	.62					
F/R	This way of assessing fits well with the social work profession	.76					
F/R	The result (output) that I had to produce in this assessment is part of the social work job					.57	
F/R	The output that was evaluated in this assessment is different from what is being evaluated in practice					.85	
F/R	The result that I had to produce in this assessment is something that a real social worker also had to produce in practice						.68
C	The criteria resembled the criteria that I have to meet in practice					.70	
C	The criteria that I had to meet in this assessment resembled the criteria used in practice					.72	
C	In this assessment I was evaluated on criteria important for the social work profession						.72
C	In this assessment I was evaluated on thing that I never have to use in real professional practice						.55
CT	The criteria that I had to meet in this assessment were clear enough				.61		
CT	Before I started the assessment it was clear to me what was expected of me				.65		
CT	It was hard to find out what was expected of me in this assessment				.82		
KL	In this assessment, both knowledge and professional skills were important	.75					

Note. O = Overall authenticity; T= Task; PC = Physical context; F/R = Form/Result; C = Criteria; CT = Criterion transparency; KL = Kind of Learning

The comprehensibility argument (Dunteman, 1989) corroborated selecting these six factors, because they were readily interpretable in the eyes of the three reviewers. Then, a new factor analysis was conducted on the remaining items that primarily loaded on these six factors. Table 3.3 shows the results of the final factor analysis. These six factors accounted for 63 % of the variance.

A closer look, from a more qualitative point of view, at the distribution of the items over the remaining factors and the items that fell out of the final factor analysis showed an interesting pattern. First, almost all items of the task, physical context, result/form and criteria dimension fell in the final factors. Second, none of the social context items loaded on the final factors. Third, the task items clustered in Factor 3 and the physical context items clustered in Factor 2. On the other hand, the result/form items and the criterion items did not cluster in the expected way. The original result/form scale contained three result items and three form items. The factor analysis showed that these items did not belong together, since the form items clustered in Factor 1, while the result items clustered with the criterion items in the Factors 5 and 6. Fourth, the criterion transparency scale was the *only* characterising element that was represented in the final factors (Factor 4). Fifth, three items of the overall authenticity scale loaded on the final factors, two of which clustered with the task items on Factor 3 and the other one clustered with the form items on Factor 1.

Three reviewers (1, 2, and 3) named the final six factors as follows:

1. Connection of assessment form with the profession (1, 2), assessment method (3)
2. Professional context (1, 3), perception of fidelity (2)
3. Content authenticity (1, 2, 3)
4. Clear expectations (1, 2, 3)
5. Job-relevant criteria (1, 3), job-related judgement (2)
6. Relevance of the output for the profession (2), job-related judgement (1, 3).

The internal consistency of the factors was calculated, for both the student and teacher group, to examine if the factors represented reliable constructs (Table 3.4). If the factors are internally consistent, interpreting these factors is more valid.

Table 3.4. Reliabilities of the final factors from the perception questionnaire.

Factors	Students (n = 115)	Teachers (n = 18)
Factor 1 (Form)	.76	.81
Factor 2 (Physical context)	.83	.91
Factor 3 (Task)	.79	.90
Factor 4 (Criterion transparency)	.76	.92
Factor 5 (Result/Criteria)	.68	.76
Factor 6 (Result/Criteria)	.69	.82

The results showed that the reliability of the factors was much better than the original scales in the student group, while the reliabilities remained high in the teacher group. These findings

might mean that these final six factors more adequately described the facets that determine assessment authenticity in the eyes of the students.

The second possible explanation for the unreliability of a number of questionnaire scales in the student group could be that the items in those scales were accurate for assessing the intended variables, but too difficult for students to fully understand. This could be a reasonable explanation, since the scales were reliable in the teacher group. To assess the reading difficulty of the questionnaire, Flesch-Kincaid Grade Level scores (Johnson,1998; Klare, 1963) were calculated per scale of the questionnaire. These scores are based on technical aspects of the reading material (sentence length and word length), without looking at the meaning of the words. Still, Calderón, Hays, Lui, & Morales (2005) showed that these scores can be used to test surveys at the item level, thereby giving valuable input for questionnaire development and improvement. Based on the Flesh Kincaid scores, the *minimal suggested reading age* (MSRA) per scale could be calculated by adding a value of 5 to the Flesch-Kincaid Grade level score (Klare) resulting in the formula:

$$MSRA = 0.39L + 11.8N - 15.59 + 5$$

where L stands for average number of words per sentence and N for average number of syllables per word. It turned out that the MSRA for the unreliable scales was 17.58, while the MSRA for the reliable scales was 14.85. The mean age of the student participants was 18.1.

Two extenuating circumstances need to be addressed at this point. First, Klare (1963) showed that people prefer to read below actual age and for a pupil to properly comprehend what (s)he is reading, the MSRA should at least be two years below the actual age of the pupil. Our findings are in agreement with this. Second, age alone does not give enough information, since not all students of the same age have the same intellectual capacities and as a result different reading levels can be expected from them. For example, even though VET students and pre-university students are of the same age, it is likely that their reading levels differ. VET is a form of vocational education that does not allow entry to further higher academic education (i.e., university) and is primarily populated by students who are work-oriented and/or not capable of successfully following an academic, pre-university, curriculum. Pre-university education is theoretical type of education and prepares pupils for university (Eurydice, 2004). To make a statement about whether or not the questionnaire scales were too difficult for VET students to understand, insight in the reading level that is normally expected from VET students, compared to pre-university students, is needed. A thorough search of literature and institutions in the Netherlands (e.g., Educational Council, Inspectorate of Education, Ministry of Education, Culture and Science, etc.), however, did not turn up data on differences between reading levels of VET students and pre-university students. To compensate for this, we gathered a representative sample of VET study material, pre-university study material, and adult education, university level course materials. For each category of material, four samples were taken and then the MSRA was calculated (see Table 3.5). This resulted in an absolute and relative norm for the suggested reading age for VET students, which was compared to the mean suggested reading ages of the questionnaire scales.

Table 3.5 Mean minimal suggested reading age (MSRA) for three educational levels.

	Number of samples	Mean MSRA
VET study material	4	15.41
Pre-university study material	4	18.10
Adult education, university study material	4	21.49

Note. The means of the three kinds of study materials are based on four examples of study material from different disciplines each.

Table 3.5 shows that the mean suggested reading age for the sample of VET study material is 15.41, for pre-university material 18.1; and adult education,university level course materials had an MSRA of 21.49. In other words, the reliable scales required a reading age that is normal for VET students, while the unreliable scales had an MSRA at a pre-university level.

Conclusion and Discussion

The main goal of this study was to examine if the theoretical dimensions of authenticity, as described in the 5DF, were recognised and corroborated in practice. The main conclusion could be that authentic assessment is indeed perceived as a multidimensional concept, but that some reconceptualisations are in order. Moreover, teachers and students seem to differ in how they perceive assessment authenticity. More specifically, teachers distinguish all the dimensions as well as their characterising elements as described in the theoretical 5DF, while students only differentiate four of the five dimensions and do not differentiate at the more specific level of the characterising element. These findings have implications for the 5DF, for future use of the questionnaire, for practice and for future research.

Students versus Teachers

Teachers recognised both the dimensions and their characterising elements as facets that determine assessment authenticity. In the eyes of the students, four of the five dimensions (Task, Physical Context, Form and Result/Criteria) determine authenticity, while students do not perceive the Social Context as a dimension of assessment authenticity, neither do they distinguish the several characterising elements. Two possibilities were examined to explain these findings. First, students perceive authenticity differently than the 5DF (and thus, the questionnaire) proposes which is reflected in new factor structure. Second, the questionnaire scales were too difficult for students to understand.

The factor structure found in the student group suggests that students have a much less elaborate perception of assessment authenticity than the 5DF proposes, while teachers seem to support the more elaborate conceptualisation of authenticity. This might be explained by the fact that students are "consumers" of the assessment, while teachers are the developers and in addition have much more experience with assessment practices and development. As a result, teachers are likely to have given assessments and the ideas behind assessments much more thought than students have. Teachers also have much more educational and practical experience, which might have changed their perception of assessment authenticity compared to students,

who might not even be aware of the existence of some characteristics (Honebein et al., 1993). Because of this increased degree of experience, teachers are likely to have more developed schemata for thinking about assessments (Sternberg, 1999). These results support the idea of Honebein and colleagues that having more practical and/or educational experience changes how one thinks of assessment authenticity.

The results of the readability analyses showed that the scales that were above the normally expected reading level of students, were identical to the unreliable scales. This might have caused the differences between the recognition of elements of authenticity of students on the one hand, and teachers and the 5DF on the other hand. The readability data also showed that teachers appear to have difficulties placing themselves in the position of students, as can be seen by the fact that a number of VET teachers rated the questionnaire scales as clear and understandable for VET-students prior to the administration of the questionnaire.

The differences between students and teachers suggest that we, as educators and instructional designers, should not automatically assume that students see an assessment as teachers see it. This stresses the relevance of investigating student perceptions of assessment characteristics as these are argued to be the motor behind student learning (e.g., Boud, 1995; Scouller, 1997).

Implications for the 5DF

The results of the factor analysis suggest that students structure authenticity partially different from what the 5DF proposes. Even though drawing firm conclusions would be inappropriate, because the number of participants (115) was relatively small for a factor analysis and it was only conducted within one student group, we think that this study does point out several interesting indications concerning the dimensions of authentic assessment most of which are corroborated by other research.

First of all, it supports the theoretical idea behind the 5DF that authenticity is multidimensional (Gulikers et al., 2004), meaning that the authenticity of an assessment depends on several assessment characteristics instead of on an overall resemblance between the assessment and professional practice.

Second, the Task (Factor 3) and the Physical Context (Factor 2) were perceived as two dimensions of authenticity. This distinction has also been made in previous theoretical research (e.g., Cumming & Maxwell, 1999) and empirical research already pointed at the individual impact of both these elements on student learning and motivation (Gulikers et al., 2005).

Third, the Result/Form and the Criterion dimensions showed a deviating pattern. Result and form were not perceived of as one dimension. The clustering of items and the interpretation of the reviewers argued for a separate Form dimension (Factor 1). The Result items clustered with the Criterion items in Factors 5 and 6. When looking at these factors semantically and at the interpretations of the reviewers, the following possible interpretation could be given: Both factors referred to job-relevant judgement, but Factor 5 refers to judgement in professional practice in more general terms, while Factor 6 refers to judgement in the social work profession in specific. A reasonable interpretation could be that these two factors can be combined in a Criterion/Result dimension that focuses on judging students on job-relevant aspect. As a check,

the internal consistency was calculated over the six items of Factor 5 and 6 combined, which was α = .68 (students) and α = .77 (teachers). This makes the interpretation of combining factor 5 and 6 into one Criterion/Result dimension more plausible.

Fourth, students did not recognise the social context as a dimension of authenticity. They apparently do not (yet) realise that much work involves social activities. The traditional school situation tends to isolate academic (individual, on-task) skills from social (group, off-task) skills, which feeds the belief that schoolwork is individual. This is corroborated by the qualitative study described in chapter 2 that examined teacher and student perceptions of the five dimensions of the 5DF. This study showed that students as well as teachers perceived the social context as the least important dimension of authentic assessment. The reason for this was not that they did not think the social context was unimportant, but that they have the strong belief, based on many assessment experiences (Boud, 1995; Samuelowicz & Bain, 2002), that assessment is an individual affair. This leaves us with three possibilities: (a) the social context is not perceived by students as one of the dimensions of authenticity and should therefore be deleted from the 5DF; (b) the social context might still be a dimension of authenticity, but the items of the questionnaire were too difficult for students to result in an internally consistent scale. This explanation is supported by the high MSRA for the unreliable scales; or (c) the strong belief that assessment by definition is individual should be changed first before students will ever be able to perceive the social context as a dimension of authentic assessment. Because teachers in this study recognised the social context, the literature study in chapter 2 supports the social context as a dimension of authenticity and the qualitative findings reported on in chapter 2 suggested that both teachers and students recognised that professional practice involves both individual and collaborative work, we argue that the social context should be kept into the framework for assessment authenticity. However, the Social Context, or the questionnaire used to examine perceptions of the social context, should be reconsidered in future research

Fifth, as said before, students perceived authenticity at a dimensional level, while teachers recognised several characterising elements within these dimensions. This suggests that the 5DF appropriately describes several assessment characteristics that are important for assessment authenticity. However, we should be aware that to influence student perceptions of authenticity, the 5DF should be used at the dimensional level.

In short, examining the theoretically hypothesised dimensions of authenticity from a practical angle suggests a new conceptualisation of the dimensions of assessment authenticity that involves a Form (Factor 1), a Physical context (Factor 2), a Task (Factor 3) and a Criterion/Result dimension (Factor 5 and 6). These factors are corroborated by both students and teachers as seen in the reliability scales in Table 3.4 and are therefore viewed as crucial elements of assessment authenticity. The reconceptualised 5DF is shows in Figure 3.1.

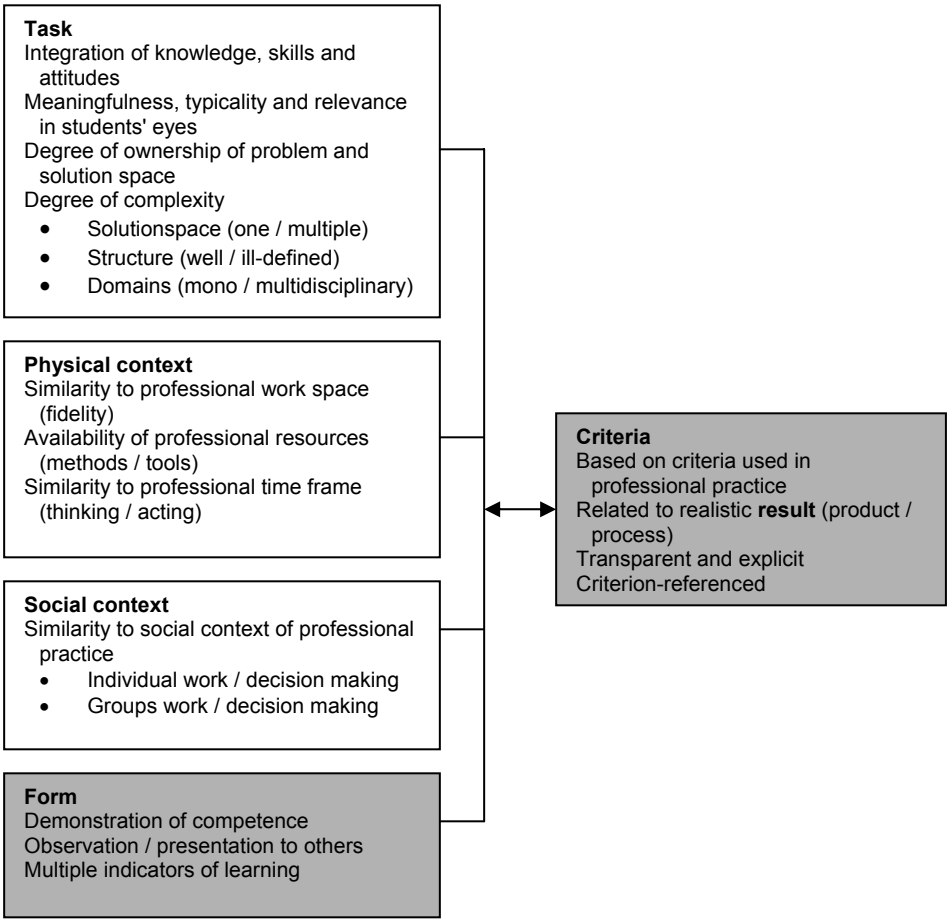


Figure 3.1. The revised five-dimensional framework for assessment authenticity. *Note.* The shaded areas in the figure indicate the dimensions in which (small) changes have taken place based on this study, compared to the 5DF that was based only on literature study (see Figure 2.2).

Practical Implications

Two problem need to be addressed with respect to the development and use of authentic assessments in educational practice (Cummings & Maxwell, 1999; Cooper, 1994; Roelofs & Terwel, 1999). First, teachers develop authentic assessments without thinking through and explicating what this authenticity means and how it is operationalised in the assessment, and second, assessments that teachers developed to be authentic are not automatically perceived as being authentic by students.

The results of this study have two practical implications that help dealing with these problems. First, the 5DF is a helpful tool for teachers or educational developers to talk and think about authentic assessments, to make implicit beliefs about authentic assessment explicit, and to

develop various kinds of authentic assessment. Second, if we want to influence student learning with authentic assessments, we have to change one or more of the following assessment dimensions: the task, physical context, the assessment form, or the result and criteria of the assessment. Instead of "dressing up" traditional assessments with some superficial real world elements (Cummings & Maxwell, 1999), making assessments authentic in the eye of the student requires changing more fundamental aspects (i.e., task, physical context, form, or criteria) of an assessment.

Future Research

To increase the generalisability of the findings of this study and strengthen the practical implications, future research should evaluate if the results found in this study also hold in other student groups in vocational education. It should be examined if the factor structure found in this study is corroborated in other groups. A new perception questionnaire for evaluating assessment authenticity should be developed and tested. This questionnaire should focus on the dimensions of authenticity reflected in the scales found in this study, as these seem to determine student perceptions of authenticity, and a new social context scale that meets the reading level of the students should be developed.

Future research should examine how student perception of the authenticity of the assessment dimensions can be increased. What makes an assessment look more like professional practice in the eye of the student? This requires an insight into what students think professional practice looks like. A student's beliefs about what professional practice is, in turn, is likely to function as a frame of reference that guides this student's perception of the authenticity of a newly encountered assessment. Furthermore, the influence is of increased student perception of authenticity on their learning should be examined. Are students stimulated to deeper learning or development of professional skills or competencies when they perceive an assessment as being more authentic? Additionally, whether different kinds of authentic assessment in vocational education have different effects on student learning and development of professional skills should also be examined. Previous studies (Gulikers et al., 2004; Gulikers et al., 2005) suggested that the various dimensions of authenticity might be of differing importance. For example, the assessment task seemed to be more important than the physical context. This raises questions such as: Is increasing the authenticity of the task enough to stimulate students to deeper learning or should more than one dimension resemble professional practice?

Besides these content related directions for future research, this study gives rise to suggestions for future research methodology. Future research that uses questionnaires, at least with students at an educational level that is comparable to VET, must consider taking student reading levels into account in the development of new questionnaires or translation of existing questionnaires. Furthermore, complementing quantitative data with qualitative data might give more insight into student, and also teacher, perceptions of assessment authenticity.

In conclusion, this study showed that exploring assessment authenticity from a practical viewpoint, by examining the perceptions of the users, has additional value over only a theoretical examination of assessment authenticity. This corroborates the idea that authenticity is, at least partly, subjective.

Chapter 4

Getting the Whole Picture: Student, Teacher and Practitioner Beliefs about Authentic Assessment[3]

This study described in this chapter examined the beliefs about authentic assessment for assessing job performance of students, teachers and practitioners, taking into account that authenticity is a multidimensional concept. Stakeholders might differ in their beliefs about what authentic assessment is, because they have had different prior experiences with assessment and professional practice. Focus groups, individual interviews, and a questionnaire were used to examine the beliefs of the assessment stakeholders, and especially the differences and similarities in their beliefs, in the field of nursing. With respect to the authenticity of the assessment task and of the physical context, the three groups mainly agreed. Differences, however, were found between teachers on the one hand and students and practitioners on the other with respect to their beliefs about the authenticity of the social context, the assessment form and the assessment criteria. In general, students and practitioners argued that teachers' beliefs were outdated and that they focused too much on technical skills at the expense of generic skills. With respect to the *social context*, assessment *form* and assessment *criteria*, teachers emphasised individual assessment, one-shot assessment, and assessment based on nationally formulated criteria that remain the same over assessments. Students and practitioners believed that authentic assessment should allow collaborative assessments, a combination of multiple assessment methods and moments, and criteria adapted to practice. Different stakeholder beliefs signal important issues in authentic assessments that need to be considered in the development of authentic assessments appropriate for assessing nursing performance.

[3] This chapter is based on Gulikers, J. T. M., Kester, L., Kirschner, P. A., & Bastiaens, Th. J. (2006). *Getting the whole picture: Student, teacher and practitioner beliefs about authentic assessment.* Manuscript submitted for publication.

Jobs and job functions have changed in the last decades and different "things" are required from employees to be successful (e.g., Birenbaum, 1996; Onstenk, 1997; Tillema, Kessels, & Meijers, 2000). 'Creative responses to the unpredictable', 'confidently taking appropriate action in unfamiliar and changing circumstances' and 'adaptive competence' are all examples of crucial features for successful job performance in the rapidly changing world (Lizzio & Wilson, 2004b). In the field of nursing, change is noticeable as well. Nursing has become more than knowing how to follow prescribed procedures and carrying out routine activities on a doctor's request. Attitudes with respect to social services have changed as well with patients being more independent and nurses being expected to deliver tailor-made care and function as responsible and independent member of a team of caretakers. They have to be able to flexibly and independently adapt their knowledge and skills to every individual patient and care situation. There is more focus on generic transferable skills and attitudes as communication, interaction, and reflection (Boud, 1998; Birenbaum, 2003; Onstenk; Tillema et al.; Wilson, Lizzio, & Ramsden, 1997). In other words, nurses not only need to be knowledgeable, they must, above all, be competent (Miller, 1990).

In order to prepare students for a labour market that requires competent employees, Vocational Education and Training (VET) programs in the Netherlands started changing from curricula focusing mostly on content knowledge, to curricula focusing on competencies needed for successful job performance (Tillema et al., 2000). For every vocational field, competencies are described at a national level, which are based on important activities of that field and which describe an integrative whole of knowledge, skills and attitudes. For nursing an example is that "a nurse should be able to adequately draw up a nursing plan considering nursing care, advising and coaching, and prevention, for the benefit of a specific patient within in certain context" (Knowledge Centre for Vocational Education and Business, 2004). By describing a profession in this way and requiring schools to develop curricula based on these descriptions it is tried to enhance the link between educational programs and work performance and to bridge the gap between learning and working.

Changing towards competency-based education not only involves changes in instruction, but changes in assessment practices as well (Biggs, 1996; Birenbaum, 1996). The assessment practices are the focus of this chapter. Tillema and colleagues (2000) suggest that in competency-based education it is not the content that is the central issue, but the assessment of the acquired competence. How can a student prove that he or she is capable of successful performance in the world of work? How can a student prove that he or she is a competent nurse, ready for the labour market? Messick (1994), Dochy (2001) and Segers (2004) argue that assessments of students' capability of successful job performance should be authentic. More specifically, to come to a valid conclusion about a student's professional functioning, assessment should appropriately reflect the professional situations and processes representative of that professional domain (Gielen, Dochy, & Dierick, 2003; Messick). In other words, assessment should be authentic with respect to students' present or future professional practice (Gulikers, Bastiaens, & Kirschner, 2004; Messick). But the question is: "What is authentic?" Or more specifically, "What elements make an assessment authentic with respect to job performance?".

Chapters 2 and 3 suggested that authenticity is multidimensional. The five-dimensional framework (5DF; see Figure 3.1) of assessment authenticity argues that one cannot simply say that an assessment is authentic or not, but that the authenticity of an assessment depends on five dimensions of that assessment that can resemble professional practice to a certain degree. These five dimensions are the assessment task, the physical context in which the assessment takes place, the social context of the assessment, the form of the assessment or the assessment method, and the criteria that are used to judge assessment performance. They can be described in more detail by characterising elements (see Figure 3.1). Based on the 5DF, it is argued that examining assessment authenticity requires looking at these five dimensions.

In addition, authenticity is subjective (Honebein, Duffy, & Fishman, 1993; Petraglia, 1998). With respect to authentic assessment beliefs this might imply that different people (e.g., doctors, patients, nursing school teachers, etc.) with different ideas about what professional functioning entails have different *beliefs* about what an authentic assessment involves. According to Samuelowicz and Bain (2002) beliefs are the characteristic way in which an individual interprets and values a phenomenon. This suggests that the beliefs that a person has about authentic assessment influence how he/she perceives or interprets the authenticity of a newly encountered assessment (Honebein, Duffy, & Fishman, 1993). Chapter 3 indicated that students and teachers *perceived* different elements as determinant for the authenticity of a certain assessment. The study reported on in this chapter goes deeper into the question of what influences these perceptions. It deals with the beliefs different stakeholders hold about authentic assessment. The question is what they believe to be important elements for an authentic assessment that is appropriate for assessing job performance. With respect to authentic assessment in VET programs, three stakeholder groups are involved, namely (1) teachers who are mainly responsible for the development and implementation of assessments; (2) students who are assessed to determine whether they are professionally competent, and (3) practitioners who are actively involved in the development of assessments to help bridge the gap between learning in school and working in professional practice. To get a complete picture of authentic assessment, this study examines what authentic assessment should look like from the viewpoint of these three stakeholder groups, taking into account that authenticity is multidimensional. The differences and similarities together will give a deeper insight in the important elements of authentic assessments.

This chapter starts by describing what determines a person's beliefs about authentic assessment and why these beliefs might be different for various stakeholders. Then, the beliefs of the three stakeholder groups are examined in the context of nursing education. Focus group interviews, individual interviews and a questionnaire based on the five-dimensional framework for assessment authenticity are used to examine authenticity beliefs from a multidimensional perspective.

Beliefs are the Result of Previous Experiences

Beliefs are the result of previous experiences (Birenbaum , 2003; Kalat, 1995; Sternberg, 2005). Especially emotionally loaded experiences, good or bad, shape our current beliefs about a

phenomenon. For example, a negative first experience with your boss makes you believe that he or she is not very personable. Or receiving a compliment when wearing a new coat makes you believe that this coat is beautiful, and receiving several compliments for this same coat, makes you never want to stop wearing it. Even though these are simple examples, the same process holds for beliefs about assessment. When talking about beliefs about assessment authenticity as a way of assessing whether a student is capable of successful job performance, two kinds of previous experiences need to be considered, namely (1) kind of assessment experiences and (2) kind and amount of experience in professional practice.

During schooling, students are confronted with different kinds of assessments that all send out a message about what assessment looks like, what it measures and what kind of preparation is needed for it (Boud, 1995; Samuelowicz & Bain, 2002; Van Rossum, Deijkers, & Hamer, 1985). These past experiences, in turn, influence a person's beliefs about future assessments. The effects of previous experiences with assessments on future beliefs about assessment and on a student's future learning for certain types of assessment are referred to as "post-assessment effects" (Gielen, Dochy, Dierick, 2003, p. 44). Until recently, most of us experienced primarily *traditional* tests such as multiple-choice tests. These tests are delivered in a written format, have to be taken by every individual student in a school setting, mostly focus on recall or reproduction of factual knowledge, and require memorisation as preparation strategy (Entwislte & Ramsden, 1983; Scouller 1997; 1998). When people have mostly experienced this kind of assessment, they will believe that this is what an assessment should look like and believe that adopting a surface study strategies, such as memorisation, is appropriate for success. These experiences, in turn, will influence how people deal with assessments in new situations and how they will learn and prepare for assessments ("long-term post-assessment effect" Segers, 2004, p. 30). These previous assessment experiences, and the assessment beliefs that are the result of these experiences, will create a reference frame for thinking about *authentic assessment* as well.

In addition, beliefs about authentic assessment are influenced by the kind and amount of experiences in professional practice. Every practice situation, even within the same domain, involves specific circumstances and communicates certain demands, norms and values. These colour a person's belief about what professional practice looks like and what kind of skills are required and valued (Lizzio & Wilson, 2004a; Radzinsky, Bouillion, Lento, & Gomes, 2001). Pena (1997) showed that when students have not yet experienced professional practice, they develop unrealistic beliefs about working and working roles based on their unrealistic ideas of professional practice. Realistic or not, these beliefs function as the frame of reference that is used to interpret authentic school activities and to determine their relevance for professional practice (Honebein, Duffy, & Fishman, 1993; Lizzio & Wilson, 2004a). Thus, the kind and amount of practical experience seems to influence what people believe is authentic.

Beliefs of Different Stakeholders Are Not Necessarily the Same

Since previous experiences influence current beliefs, students, teachers and practitioners could be expected to differ in their beliefs about authentic assessment because they will differ in their assessment experiences, especially since teachers and/or practitioners are now the assessors and

students the assesses, and definitely differ in their experience in professional practice. With respect to assessment practices, most adults (i.e., practitioners and teachers) will only have experienced traditional forms of assessments as multiple-choice tests or short-answer tests (Segers, 2004; Van Rossum & Hamer, 2003). However, in the last decade many educational changes have taken place and many schools have had to deal with new ways of instruction and assessment (e.g., problem-based learning with group-work assessments, life-long learning with portfolio, e-learning with e-assessment, etc.). As a result, current students might have a broader assessment experience than adults. Teachers, however, might have been confronted with changing the assessment practices, thereby broadening their experiences with assessments compared to practitioners.

With respect to practical experiences, the three groups naturally differ. In regular VET programs in the Netherlands such as nursing, students usually have no experience in professional nursing practice when they start studying. Still, these students have chosen to become a nurse based on their idea of what being a nurse means (e.g., "I want to work with children in a hospital"). During their studies, they immediately start doing their internships in hospitals, mental institutes, home care or geriatrics, which changes their ideas of the nursing profession. It can broaden their idea of the profession and as a result make it even more interesting for them, while on the other hand, they may decide to quit, because they experience that the beliefs with which they started their studies are not in line with professional practice (e.g., "now I am washing old people all day long, that was not my idea of being a nurse"). Practitioners and teachers both have experience in professional practice, but while practitioners still have both feet in the real world of the profession, teachers often are no longer working in this professional context. This means that practitioners, while having more up-to-date knowledge and experiences about professional practice, and thus beliefs about it, are attuned to only one setting, which might make their beliefs about the nursing profession less broadly compared to teachers. Teachers, on the other hand, might have dated ideas or a more academic or clinical view of the profession.

Several studies underlined that different stakeholders differ in their beliefs about authenticity or professional practice. Gulikers, Bastiaens and Martens (2005) and Roelofs and Terwel (1999) showed that teachers' and students' ideas of authenticity possibly differed because they had different beliefs about professional practice. Teachers believed that their educational practices reflected 'the real world', while students had different beliefs about this real world and as a result experienced them as less authentic. Teachers and practitioners were found to emphasise different attributes as most important for professional life, which influenced their beliefs about authentic activities in education (Lizzio & Wilson, 2004a; Radzinsky et al., 2001). While academics felt that authentic practices need to focus on operational competences, practitioners paid much more attention to personal attributes and interpersonal skills.

All of this lends credence to the premise that since authenticity is subjective it is crucial to make the beliefs of different stakeholders explicit for authentic practices to become successful and beneficial for learning. Addressing beliefs of involved parties is argued to be crucial for

educational innovations, especially when they oppose current beliefs (Gibbs, 1992; Radinsky et al., 2001; Van Rossum & Hamer, 2003).

This study examines the multidimensional construct of authenticity from the viewpoint of the three stakeholders in authentic assessment practices in VET education. The research questions of this study are:

1. What are the beliefs of students, teachers and practitioners in the field of nursing education about the dimensions of authentic assessment as described by the five-dimensional framework (5DF)
2. What are the similarities and differences in the beliefs of these three groups.

The differences and similarities between the three stakeholder groups concerning the facets of authenticity point to important elements that need to be taken into account when developing authentic assessments

Method

Participants

First year students (n = 34) and their teachers (n = 12) from a VET college for nursing participated in this study. In VET in the Netherlands, students start doing internships from the start of their studies, mostly for one day a week. Thus, these students had already gained some experience working in professional practice. Moreover, the students and teachers had experience with both school and workplace assessments. In addition, five practitioners working at various health care institutes (mental hospital, general hospital, geriatric care) participated in this study. All these practitioners were also mentors of nursing internships.

Data Collection

Semi-structured focus group interviews, individual interviews, and a questionnaire were used to gather data. Focus groups were held with students and teachers and individual interviews were conducted with practitioners. A random selection of the students and teachers participated in the interviews. Three student focus groups (n = 4, n = 6, n = 8), one teacher focus group (n = 4) were used. The five practitioners were interviewed individually. During the semi-structured interviews the respondents were given the opportunity to freely express all their ideas, opinions, and attitudes concerning what elements they felt are important in authentic assessment. The dimensions of the 5DF (Figure 3.1) were used as interview stimuli, but participants were also stimulated to come up with other topics that they believed were important for authentic assessment. All interviews were audiotaped.

To gain more information about the beliefs of the different stakeholders and to increase the reliability of the study, the qualitative data were complemented with a quantitative 17-item beliefs questionnaire filled in by teachers (n = 12) and students (n = 34). The reliability for the beliefs questionnaire was α = .77 for students and α = .76 for teachers. The quantitative data were used to complement the qualitative data from the interviews. Just as the interview questions, the questions in the questionnaire were designed using the 5DF (Gulikers et al., 2004) for assessment authenticity as a guideline. Pilot testing of the questionnaire showed that (a) a frame of reference

for interpreting the questions was required; instead of asking "an assessment should…" it was important to say "a competency-based assessment that focuses on assessing professional functioning should…", and (b) the answer categories should be semantic differentials with pairs of statements about competency assessment that were contradictory to each other. An example is "A competency assessment that focuses on assessing professional functioning can best take place A) at school B) in practice". Participants had to point out on a 5-point scale which statement they preferred, with the score "3" expressing no preference for A or B. The questionnaires were the same for both students and teachers with the only difference that the words "I", "me" and "mine" in the student questionnaire were replaced by "the student", "he/she" and "his/her" in the teacher questionnaire.

Analysis

The interviews were analysed as described in Baarda, de Goede, and Teunissen (2001). First, the audiotaped interviews were transcribed and irrelevant information was filtered from the interviews. Information was defined as irrelevant when it had nothing to do with assessment. The remaining text was parsed into fragments. A new question always signalled a new fragment. Within one answer, fragments were parsed according to topics discussed; one fragment contained one topic. After this, the fragments were coded first by using a coding scheme based on the five dimensions and their characterising elements (see Figure 3.1). Fragments that could not be coded as belonging to one of the five dimensions, received the label *other*. This necessitated the development of new codes reflecting the main themes of the *other* fragments. Then, a summary was made of each interview, reflecting the beliefs about the dimensions of authenticity based on the 5DF and additional themes. Finally, a qualitative description per group (students, teachers and professional nurses) was made, as well as a scheme, based on the 5DF, to compare the beliefs of the three groups.

To increase the reliability of the data analysis, the summaries of the practitioner interviews and the teacher interviews were sent to the participants to check the accuracy and completeness of the interpretations ("member checking procedure"; Guba & Lincoln, 1989). To minimise the influence of personal interpretation of the data, the interviews were fragmented, labelled and summarised independently by two researchers. After that, the interpretation of the interviews rested upon careful reflection and discussion between two researchers, one of whom was not involved in conducting the interviews.

The 17 items in the questionnaire were analysed per item to indicate the importance of different assessment facets for authentic assessment in the eyes of students or teachers. The items were also combined in an authenticity scale. This scale gave an overall indication about how important participants felt it was that an authentic assessment should reflect real professional practice. Student and teacher ratings were compared to examine whether they differed in how important they believed authenticity to be for competency assessment.

Results

The results from the interviews are described per stakeholder group and per dimension of the 5DF. Quotations that accurately depicted the discussions are provided. Numbers indicating

fragment and focus group number are given in parentheses. After describing the three stakeholder beliefs concerning a dimension, the similarities and differences between the three stakeholders are summarised. Table 4.1 summarises the findings for the three groups per dimension and shows the differences and similarities between the three groups. The final section describes the results of the questionnaire.

Task

Students. The task should be based on *professional competencies* or *critical professional activities* instead of being based on school activities, course books or specific activities at a specific workplace. The description of the task should be broad enough for every individual student to contextualize the task to the work setting and interest, "if the task gets too specific it is impossible to adapt it to your own work (26, 3)". Students want some *ownership* over the content (e.g., which patient, with what kind of problem) and the *complexity* level. This makes the task *relevant and meaningful* and gives student the opportunity to learn, "if you choose an easy task, you are not challenging yourself (98, 3)". However, every task description should be structured with minimal requirements to make the tasks as comparable as possible between students and that schools should not try to make the task more complex than necessary, because this makes the task less authentic, "my school asks me to first read the whole file of the patient. But I know the patient already, so then it is stupid to require me to read the whole file again (120, 1)".

Teachers. The task should be based on *professional competencies* and be described in terms of *performances of the nursing profession*. The task should entail a broad description, which the student, in collaboration with a mentor at the workplace, can translate to the work setting, "the description should be applicable in every nursing context … it has to do with translating (92-94, 4)". By giving students some *ownership* over the translation of the task to suit their individual context and interest, the task becomes *personally relevant*. Students also have *ownership* of the time of performance, "the student has to indicate when he/she is ready for the assessment (112, 4)". The *complexity* level should be based on what is required in the workplace, but guidelines should be set in the task description to prevent the task to become too complex for first year students.

Practitioners. Tasks should be representative for the nursing profession and should be based on *up-to-date competencies* and *main activities* required in the profession. The activities of a nurse have changed and assessment tasks should take change into account:

> In the past, another person responsible (e.g., the doctor) controlled everything and you only carried out orders. Nowadays, you are expected to think along and help coordinate. I think it is important to prepare students to do this when they graduate (113, 5). A problem is that teachers often do not have up-to-date experience in practice, which is reflected in assessment practices. Teachers have worked in practice once, but they already left a long time ago … they don't take into account that situations change and that the way of thinking changed in the health care. They hold on to the past and how things are described in books (114-115, 5).

The task should be a broad description based on *knowledge, skills and attitudes*, which students should give meaning in their own work context. This makes the task relevant for every

individual student. Students should be given much *ownership*; they are mainly responsible for interpretation of the task and they should be given space to fill in the assessment tasks with problems they encounter in practice. Teachers and practitioners have a supporting role in defining the specific task. Tasks should resemble the *complexity* level as required by practice. Tasks should not be made easier for students.

Summary. The three groups agreed that the task should be described in terms of professional competencies or activities and not specified to a specific work setting or school activity. Also, to make the assessment personally relevant, students should have some ownership over the content of the assessment, which gives them the opportunity to translate a broad task description to their own work setting and interest. In other words, a task should be transferable to various contexts. However, both practitioners and students complained about a decrease in authenticity of an assessment tasks when teachers are not up-to-date with respect to professional competencies or require irrelevant or redundant actions from students.

Physical Context

Students. Summative assessment of professional functioning should be done *in practice*. The difference in real practice is that unexpected things can happen and an assessment should test if a student is able to deal with this, "you have to be able to play along with the patient, this is what makes it more than 'doing the trick' (131, 1)". However, practicing in school is crucial.

Teachers. Summative assessment of professional functioning should be done *in practice*, "it is difficult to set up these things in a simulation. It might be possible, but ask students, it is a bit fake (37, 4)". However, practicing in school is crucial to prepare students to deal with real life situations. First a lot of theory needs to be mastered and then students have to practice in different situations and in different ways:

> First students have to master a lot of theory, than they have to practice with simulations of patients in all different kinds of situations. Cases from practice, question people who work in practice about what can happen in practice and based on these experiences go into practice themselves (8, 4).

Practitioners. Summative assessment of professional functioning should be done *in practice* in order to test if students are able to deal with the unexpectedness of reality. However, an assessment will never be completely the same as real life, because a student will never have final responsibility while a professional nurse will. School is a good place for practicing and learning and for taking away insecurities. Theory testing, simulations, role-plays, etcetera can be done in school and used as assessments, but they should always be combined with final assessments in the workplace. Generic skills and attitudes are very difficult to test in school.

Summary. The three stakeholders agreed that summative authentic assessment should take place in a work setting, while practicing with whole tasks (i.e., based on the same competencies as the summative assessments) in school situations was crucial.

Table 4.1 Student, teacher, and practitioner beliefs about authentic assessment

Dimension	Students	Teachers	Practitioners
Task			
Content	Broad description based on professional competencies and professional activities instead of specific work setting (e.g., home care) school activity or course book	Broad description based on professional competencies in terms of professional performances instead of focused on a specific work setting	Broad description based on up-to-date professional competencies defined by knowledge, skills and attitudes applied in context. Experiences and questions encountered in practice should be used to define assessment tasks
Relevance	Translating broad description to own workplace	Translating broad description to own workplace	Translating broad description to own workplace
Ownership	Over content and over complexity level	Over content and time	Over content. Student is mainly responsible for the content
Complexity	Responsibility of the student and based on what is required in practice. School should not make assessment more complex than necessary	Complexity level that is required in practice, but school has to set boundaries to complexity	Complexity level that is required in practice, assessment should not be made easier
Student level	Study year should be taken into account	Study year should be taken into account	Study year should be taken into account
Structure	School has to set minimal requirements to make assessments comparable between students		Teacher and mentor have supporting roles in describing the task.

Dimension	Students	Teachers	Practitioners
Physical context			
	Summative assessment in workplace because of unexpectedness factor, practicing in school is crucial	Summative assessment in workplace, practicing in school is crucial	Summative assessment in workplace because of unexpectedness factor, practicing in school is crucial
Social context	Mostly individual, but when task asks for it, collaboration should be allowed Collaboration and giving feedback important for authentic assessment and in this case collaborative assessment is necessary	Individual since real life is individual as well	Asking for assistance should always be possible. For professional functioning it is important know when and how to ask for help. Collaboration is important nursing competence and should be assessed
Form			
Demonstration of competence	Summative assessment of job performance requires (1) knowledge; (2) technical skills; (3) generic skills. Knowledge can be tested in school with paper and pencil test or knowledge and technical skills can be combined in a performance assessment. Generic skills require observation while doing regular job	Summative authentic assessment focuses on performances that requires integration, adaptation and application to specific context. No separate testing of knowledge, only embedded in the performance assessment Personal development should be included	Summative assessment of job performance requires (1) knowing (translation from theory to practice) (2) technical skills; and (3) generic skills and professional attitude. Knowledge testing can be done in school and should not be responsibility of practice. However, knowledge testing as translation from theory to practice can be integrated in performance assessments. Main focus should be on generic skills, which should be observed while doing regular job

Dimension	Students	Teachers	Practitioners
Multiple indicators of learning	Summative assessment involves more moments and more methods Technical skills can be assessed in a number of fixed moments. Generic skills require long-term observation and information should be gathered from various sources Formative assessment is preferred, combined with feedback to improve job performance	Summative authentic assessment can be one-shot when (1) all educational activities are directed to summative assessment; (2) there are possibilities to practice in various situations; and (3) enough knowledge can be gained. Formative assessment is mostly for teachers to check if a student is ready for the summative assessment	Summative assessment involves more moments and more methods Technical skills can be assessed in fixed moments. Generic skills require long-term observation and information should be gathered from various sources. Generic skill development should be assessed regularly in a oral conversation with student Formative assessment with feedback is preferred to improve job performance and develop students perspective on him/herself as professional Initial formative assessment to assess if students understand and can articulate the competencies, and are able to translate them into practice
Criteria			
Developers	School and practice	School, based on national standards	School should set conditions, practitioners should be informed to align the criteria to what is really needed and valued in practice Practice should be allowed to add criteria

Dimension	Students	Teachers	Practitioners
Related to realistic result	Broad criteria for generic skills that can be specified in the work context to suit the methods, procedures and ways of working in that context, instead of being fixed to the school situation Specific and concrete criteria for technical skills Also criteria should be set for theory	Broad criteria that can be translated to every context Criteria should remain the same over assessments	Concrete criteria for technical skills and broad criteria for generic skills. Now school criteria, and as a result students, focus too much on technical skills at the expense of generic skills. Broad criteria should allow specification in the work context to suit the methods, procedures and ways of working in that context, instead of being fixed to the school situation
Transparent and explicit	School needs to be explicit towards students about what is expected. School needs to communicate to practice what practitioners and students are expected to do.	--	School should communicate to practice what the criteria are and what they mean by them in practice. Students should have a clear idea of what the criteria and competencies underneath the criteria are and they should be able to articulate them. School or practice should test if students do indeed understand what the competencies and criteria mean and how they can be translated to practice.

Social Context

Students. This dimension did not receive much attention. Students felt that they should be able to perform *individually*. However, if a situation asks for it, they should be given the opportunity to work together during the assessment, "in my case, somebody had to help me, because this patient was simply too heavy to wash on my own (67, 1)". A suggestion of students to allow collaboration in an assessment, while still being able to test the individual student, was to let the student tell the other person what to do, "I should tell my colleague what to do, then I am actually doing the performance on my own, she can see that I know what I am doing (21, 1)". Students saw collaboration and giving and receiving feedback as two important competencies for nurses. These things can only be assessed if working together is allowed in the assessment.

Teachers. Here too, there was not much attention for this dimension. Students should do the assessments alone, because in practice nurses work *individually* as well, "people do it individually in practice (127, 4)".

Practitioners. Students should always have the option to ask for assistance; in reality it works this way as well, "what is important is that students learn when to ask for help (73, 5)". Allowing students to *work together* and to ask questions is very important for their development and to get insight in what they know and can. In addition, collaboration is a very important professional competence of being a nurse and should be assessed.

Summary. All three groups did not pay much attention to the social context of the assessment. However, an important difference was found between students and practitioners on the one hand and teachers on the other. Students and practitioners argued that working together and asking for assistance should be allowed in the assessment, when the task required it. Teachers believed that also authentic assessment should be carried out individually.

Form

Students. Successful job performance requires (1) knowledge, (2) concrete or technical skills, and (3) generic skills such as communication and collaboration. All three elements should be part of a summative authentic assessment. Students felt that knowledge testing is also important for working in practice, "because you just perform better when you know the theory (28, 1)". Knowledge testing can be done in school with a separate knowledge test, but can also be integrated in a performance assessment by asking students to explain what they are doing and why, "I think that they should ask questions and continue to ask questions like 'why are you doing that', instead of only looking at your performance (3, 1)". In addition, assessing professional functioning requires *more assessment methods* and *more assessment moments*. Assessment of technical skills and knowledge can take place at a *fixed moment*, but more than once. One-shot assessment with somebody watching your actions makes students nervous which makes the performance less realistic. Generic skills cannot be tested at a fixed moment. This asks for an ongoing assessment over a longer period of time by *observing* a student during normal work. The assessment of generic skills requires input from *various sources* (e.g., colleagues), "that you also have pieces of information from colleagues, how they think you are collaborating (163, 1)". Finally, students felt that assessing job performance requires *formative assessment* moments combined with feedback, directed towards improving job performance and learning.

Oral conversations are preferred as assessment form for these formative assessments, "that you just get feedback about what you can change. Not that that has to be formally assessed or something, just a conversation about how things are going (112, 2)".

Teachers. An authentic summative assessment requires *performance of an integrated whole* of knowledge, skills and attitudes. It should stimulate students to flexibly adapt a broad task and criteria to their own situation, "the student should constantly be asking him/herself 'at what ward am I working, how can I apply this task in this situation, how can I translate this task in such a way that I show my competence?' (87, 4)". Knowledge testing should not be done separately, but should be integrated with the performance "you should ask questions to get a good idea what a student knows while or after performing, this way it should be more than only carrying out skills (47, 4)". *Personal development* should also be part of an authentic assessment, "personal development is present in every activity, because you always take yourself with you. This should be taken into account in an authentic assessment (120, 4)". A summative authentic assessment could be *one-shot* under certain conditions: (1) all educational activities should be directed towards the summative performance; (2) students should be able to practice in various situations (in school and in the work context) and receive feedback; and (3) students should be given the opportunity to gain required knowledge. *Formative assessment* is used to judge if a student is ready to do the summative assessment.

Practitioners. A summative assessment for job performance should include several elements, namely (1) knowledge, especially knowing how to translate theory into practice or knowledge into performance; knowing why and knowing the consequences of your actions. This kind of theoretical knowledge is needed for flexible job performance; (2) technical skills; and (3) generic skills, professional attitude and mentality. Generic skills have gained much importance in the nursing profession, with special attention for organisational skills, independence and responsibility, and coordinating skills, "nowadays, much more is expected from graduated nurses, for example with respect to organisational skills (78, 6)". The main focus of the assessment should be on generic skills and attitudes, because these are most important in practice, while most problematic for students, "in the assessment, little attention is paid to attitude, while these are the things that count here [in professional practice] (23, 7)" and "communication skills often make performance in practice problematic (59, 8)". Authentic assessment should be a *combination of tests.* Technical skills combined with underlying knowledge ("why") can be assessed at fixed moments. For generic skills you need to observe a student for a longer period of time, while working:

> To wash and dress a patient, purely the skills, that can be scored step-by-step, but this does not reflect how you approach the patient, the attitude that students has towards the patient or colleagues, you cannot score that in that way [step-by-step, in one moment] (60-61, 8).

For good assessment of generic skills, input from various sources (e.g., colleagues) should be gathered, "the whole team is involved in assessing how a student functions on the work floor (64, 8)". *Formative assessment* is very important in authentic assessment to help students grow and get an insight in their own development as a nurse "reflecting on yourself, that should be in such

an assessment. That does not have to be of a judgemental nature, but focused on 'what is my position and place in this work field' (31-32, 7)". A formative assessment that is suggested to start authentic assessments with is to examine whether students understand the competencies and know what they mean in practice. Oral conversations are preferred as assessment form for the formative assessments "I think that you have to think of an oral test, because it gives you a more complete overview (94, 8)".

Summary. Students and practitioners argued that job performance required a combination of assessments, could not be assessed in one moment, and required gathering information from various sources (e.g., colleagues). Authentic assessment required assessment of knowledge (main focus on knowing "why" and "how"), technical skills and generic skills. Formative assessment, in the form of oral conversations, had the purpose of helping students grow, develop and improve their job performance. Teachers, on the other hand, felt that one-shot summative assessment should be possible if formative assessments showed that students were ready to perform the summative assessment. Separate knowledge tests were not acceptable in authentic assessments.

Criteria

Students. The *school and practitioners* should both be involved in criteria development. Practitioners are important because they know exactly what is important in practice while teachers do not, "teachers only give a certain course, they don't know what practice really looks like (42, 2)". Moreover, criteria should allow *performance according to the guidelines of a specific workplace* (i.e., procedures, protocols or way of working used at this institute) instead of being restrictive to chosen school formats:

> I think that if a teacher would have come to assess me at my workplace, I would have failed, because I don't do it like school wants me to. At my workplace, I do things the way they taught me, like they do it themselves (44, 2).

For assessing job performance an assessment should involve criteria about (1) theory, where students need to be able to explain why they do the things they do; (2) concrete (stepwise) criteria for technical skills, "directly directed to the technical skill (e.g., measuring blood pressure) (152, 1)"; and (3) more general criteria for generic skills and attitudes. Finally, school should be *clear* about their expectations towards both students and practitioners.

Teachers. The school translates national standards into assessment criteria and practitioners have no responsibility in criteria development. Criteria should be general and remain the same so that students adapt them to their own context and use the same criteria in several contexts. National standard criteria are the start of the assessment and by specifying these criteria to the context and the difficulty level that is required, the assessments change all the time.

Practitioners. The school has to set conditions or boundaries for the criteria, but practitioners should be involved in developing criteria, because they know what facets are important for the nursing profession. Moreover, criteria should focus using theory and competencies in practice, instead of being theoretical descriptions, "What do they mean exactly? They [the criteria] are nicely formulated, but how can I do that in practice? Mostly they are formulated too theoretically

(135, 8)". Criteria should cover all they want to know because they are the start for filling in the rest of the assessment and they guide student learning. There should be concrete criteria for technical skills, but also criteria for generic skills and attitudes.

> Now, most [criteria] are focused on technical skills, and none on professional attitude, coordinating, delivering quality care, and so forth. It is difficult to get students to do something about this, because they only see the criteria and say 'this is what I have to accomplish, this is what I am assessed on and that's it' (43, 5).

In addition, criteria should allow performance according to the rules and procedures of a specific workplace, "protocols that we use here are not synonymous to what students learn in school … but the student should be able to use the protocol that is used in this work setting (30, 8)" and the workplace should be given the opportunity to add criteria based on what is important or required in this specific workplace. Criteria should also be very clear and transparent to let everybody (students as well as practitioners) know what is expected. This also involves that criteria should not only be given, but they should also be understood, "what you notice is that student don't understand the competencies that they have to prove. They can not articulate them and translate them to what they mean in practice (91, 8)".

Summary. Students and practitioners were more in line with each other than with teachers. They stressed the importance of involving practice in the development of criteria for two reasons. First, teachers were not always up-to-date with developments in professional practice. Second, practitioners should make sure that criteria were formulated in such a way that they were usable in practice instead of being too theoretical and that they allowed performance according to the guidelines used in a specific workplace. Teachers argued that practice had no responsibility in criteria development because assessment criteria are determined by national standards. Moreover, teachers argued that one general list of criteria should be used. This list should remain the same over assessments and needed to be adapted to different situations to make up different assessments. Students and practitioners stressed that different assessment methods should be combined with their own set of criteria for clarity purposes and to stimulate learning all the relevant aspects of job performance. In addition, students and practitioners stressed that criteria and expectations needed to be clearly communicated and understood. This should include evaluating if students and practitioners understand the theoretical criteria and competencies and understand how these can be translated to practice.

Additional Themes

Four new codes were defined to describe the themes in the *other* fragments that were related to assessment, but could not be coded with the codes of the 5DF (e.g., "during skill training courses in school can learn how you have to do something and during an internship you have to perform this" (119, 1)). Four themes came up in the interviews of the three stakeholder groups. These were "the assessor", "communication and collaboration between school and practice", "the influence of assessment on student learning", and "the alignment between instructional activities and assessment". These additional issues stress important elements of assessment practices, but are not directly connected to authenticity, because they do not refer to the relationship between

beliefs about what successful performance is and beliefs about what an assessment for job performance should look like. Therefore, these themes are not further discussed.

Quantitative Data

Table 4.2 shows the scores per item of the students and teachers.

Table 4.2. Mean and Standard Deviations for Students and Teachers on the Beliefs Questionnaire

		Students (n = 34)		Teachers (n = 12)	
		M	*SD*	*M*	*SD*
1	It is important that the test has a connection with professional practice	4.32	.95	4.67	.49
2	Task should resemble the tasks in professional practice to a great extent	4.15	.93	4.50	.91
3	It is important that the environment in which the test is performed resembles professional practice	3.50	1.24	4.83	.39
4	A competency test should be carried out together	3.68	1.25	4.00	1.21
5	A competency test should focus on whether or not I can act in a professional situation	3.74	.99	4.58	1.17
6	It is important that I should produce something that is relevant for my future work	4.59	.82	4.67	.65
7	The criteria on which I am assessed, should be the same as the criteria on which I am assessed in professional practice	4.00	1.10	4.00	128
8	It is important that the test resembles professional practice	4.26	.93	4.75	.45
9	It is important that the task that have to be carried out in the assessment are the same sort of task that have to be carried out in later profession	4.47	.75	4.64	.51
10	The best place for competency assessment is the workplace	3.47	3.99	4.64	.51
11	I should be able to ask for help during a competency assessment, if I need it	3.18	1.36	3.00	1.18
12	It is about whether or not I can use my knowledge	3.85	1.18	4.55	.93
13	It is important that I should perform actions that I must also do in professional practice	4.32	.88	4.73	.47
14	The criteria should focus on professional demands	4.12	.88	4.67	.49
15	It makes a lot of difference whether a competency assessment is done with a dummy/role-player or with a real patient	4.03	1.34	4.17	1.03
16	The actions/products I have to show, must be strongly focused on professional practice	4.29	76	4.58	.52
17	It is important that the criteria on which I am assessed, are important for later professional practice	4.50	.75	4.75	.45
	Overall	4.03	.47	4.45	.37

First, one-sample t-tests were used to investigate if the scores on the whole questionnaire and the scores on the single items differed significantly from the neutral score of 3 for both students

and teachers. This was the case for all items except for item 11, which was related to the social-context dimension. All the other item scores and the overall-score were significantly higher than 3 ($p < .05$), meaning that both students and teachers believed that authenticity of several assessment characteristics was important for competency assessment.

Independent sample t-tests were conducted to investigate whether students and teachers differed in their beliefs about authenticity on the item level and the overall scale. Students and teachers differed significantly on the overall score, $t(44) = 2.70$, $p = .01$, indicating that teachers placed more importance on authenticity in competency assessment than students did.

At the item level (see Table 4.2) the difference was significant for four items, namely item 3, $t(44) = 3.65$, $p = .001$; item 5, $t(44) = 2.43$, $p = .02$; item 10, $t(44) = 3.73$, $p = .001$, and item 14, $t(44) = 2.04$, $p = .05$. Items 3 and 10 are related to the physical context, item 5 is related to the form, and item 14 is related to the criteria of authentic assessment. In all four cases, teacher scores were higher than student scores, indicating that teachers said to value the authenticity of assessments more than students did.

Discussion

This study examined the beliefs of three stakeholder groups concerning authentic assessment. Why is studying beliefs concerning authentic assessment so important? As Figure 4.1 shows, beliefs tend to influence people's perception or interpretation of new situations, which in turn influences their behaviour (e.g., Samuelowicz & Bain, 1992; 2002; Van Rossum et al., 1985).

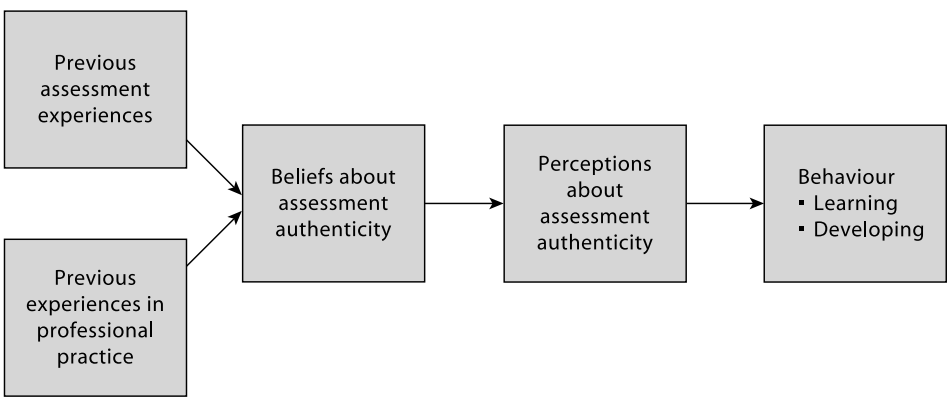


Figure 4.1. Relations between beliefs, perceptions and behaviour

Specifically, stakeholders' beliefs about authentic assessment or professional practice will influence if they perceive and value a new assessment as authentic, that is, as a good reflection of professional practice. For students, perceiving assessment as relevant for professional practice positively influences their learning behaviour and their motivation to learn (Gulikers et al., in press; Herrington & Herrington, 1998; McDowell, 1995; Lizzio & Wilson, 2004a). For teachers and practitioners, beliefs about what an assessment for job performance should look like guides their decisions in developing new authentic assessments (Petraglia, 1998; Radzinsky et al., 2001). This

study showed that students, teachers and practitioners differ in several ways in their beliefs about authentic assessment. We argue that it is crucial to make these different beliefs explicit and take them into account in the development of new authentic assessments, because of the expected relationships depicted in Figure 4.1.

One of the most important findings in this study is that, with respect to several dimensions of authentic assessment, students and practitioners are more in agreement with each other than with teachers. Students and practitioners in this study had recent experiences in professional practice, while teachers did not. Teachers' beliefs of professional practice and the assessments that they develop, are judged as not always up-to-date with developments in professional practice and as a result not always authentic in the eyes of students and practitioners. One important difference between teachers on the one hand and student and practitioners of the other hand, is that teachers tend to focus too much on assessing technical skills at the expense of generic skills, while generic skills like organising, acting responsible and coordinating are the kind of skills that are most important in the nursing practice nowadays, and other professional fields as well (Boud, 1998; Handal & Hofgaard Lycke, 2005; Semeijn, 2005; Wilson at al., 1997).

Another important difference is that, according to teachers, separate knowledge testing should be avoided, while students and practitioners argue that knowledge testing should be part of an authentic assessment. Several studies argued that a good knowledge base is crucial for professional functioning and that knowledge testing can, or maybe even should, be part of competency-based authentic assessment (Baartman, Bastiaens, Kirschner, Van der Vleuten, in press; Segers, Dochy, & DeCorte., 1998; Straetmans, Sluijsmans, Bolhuis, & Van Merriënboer, 2003; Van der Vleuten & Schuwirth, 2005).

With respect to the social context of the assessment, teachers stress individual assessment, while students and practitioners argue that professional practice requires students (and professional nurses as well) to always be able to ask for assistance when they need to. Collaborative assessment should therefore have a place in authentic assessment.

Finally, in the eyes of students, teachers tend to be too strict in their criteria meaning that they want students to perform according to the guidelines of the school, while students, supported by practitioners, argue that authentic assessment should require students to perform according to the guidelines of the specific institute they are working in.

Two considerations can be given with respect to these discrepancies. First, school managements might consider stimulating teachers to continue working in practice for a certain amount of their time. Second, in line with one of the ideas of the assessment culture (Birenbaum, 1996), authentic assessment practices should benefit from involving students as well as practitioners in the development of the assessments. This study showed that especially with respect to the form of the assessment, the combination of assessment methods and moments, possibilities to include collaborative assessments, and the development of criteria, collaboration between the three stakeholder groups might be beneficial.

Critical Remarks and Future Research

Before discussing possibilities for future research, some critical comments are in order. First, authenticity is not the only quality criterion for competency assessment (Baartman et al., in press;

Dierick & Dochy, 2001) and the additional themes found in this study also referred to other quality criteria like reliability, consequential validity and alignment with instruction that are important for good assessment practices. The importance of making an assessment more authentic depends on the goal of the assessment. This study does not want to state that all new assessments in VET education should aim at achieving high authenticity. It is argued that when the goal of the assessment is to assess job performance, authenticity is a crucial quality criterion for assessment.

A limitation of the study relates to it being a relatively small-scale study restricted to one setting, nursing, which is a well-defined profession in a non-profit sector. All participants had experience with both school and workplace assessment and the students already had some experience with working in practice. All these variables (well-defined profession, non-profit, previous experience with workplace assessment and with working in practice) might influence the results and limit the generalisation of the findings of this study. For example, the study reported on in chapter 2 explored first-year nursing students' ideas about authentic assessment and showed that their beliefs differed in several important ways from the first-year nursing students in the current study. Students in the previous study did not stress the importance of assessing in real professional practice, and with respect to increasing the authenticity of the assessment form, they mainly talked about paper-and-pencil tests using written case descriptions of realistic patients. In contrast to the students in this study, these students had no experience with assessments in practice. This supports the premise of this study that previous experiences with assessments and with professional practice influence people's beliefs about what authentic assessment should look like (see Figure 4.1). For further research it would be interesting to examine the beliefs of students, teachers and practitioners who differ in some of the aforementioned variables to find out what the influence of the variables is on their beliefs and what this means for authentic assessment practices. For example, do authentic assessments differ for different study domains? When there is more insight into the beliefs of authenticity, an additional interesting question would be whether authentic assessment practices should be adapted to fit current beliefs or if we should focus at changing those beliefs. For example, are work assessments necessary from the start or can schools offer freshman students a realistic preview of professional practice? Van Rossum and Hamer (2003) suggested that in order to give educational changes the possibility to become successful, it is crucial to make current beliefs explicit and help people develop their beliefs to fit with the new educational ideas. Thus, what is the effect of making these beliefs explicit and what should or can be done to help develop beliefs to fit with the new educational ideas?

The quantitative data from this study showed that both teachers and students feel that it is important that competency-based assessment resembles professional practice. For teachers this seemed even more important than for students. The more detailed qualitative data corroborated that different stakeholders agree that authentic assessments should resemble professional practice, but that the meaning of 'resemblance' differs between groups. This study showed that stakeholder beliefs about authentic assessment differ and previous research showed that these differences are reflected in authentic assessment practices (Gulikers et al., 2005; Radzinsky et al.,

2001; Roelofs & Terwel, 1999). This means that the authenticity of actual assessments is might be perceived differently by different stakeholders, while *student perceptions*, not teacher perceptions or intentions, determine student learning.

To conclude, even though relatively small-scale and restricted to one setting, examining and comparing the viewpoints from students, teachers and practitioners is innovative and seems to lead to a lot of useful information concerning relevant aspects of authentic assessments and possible problems, with respect to different beliefs about authenticity, that can help bridging the gap between learning and working.

Chapter 5

Authenticity is in the Eye of the Beholder: Student and Teacher Perceptions of Assessment Authenticity[4]

The study in this chapter examined if different stakeholders perceived the authenticity of an assessment differently. It examined if students and teachers differed in their perception of the authenticity of the dimensions of the five-dimensional framework. Subsequently, it investigated if freshman and senior students, who differed in their amount of practical experience, differed in their perceptions of the authenticity of the same assessment dimensions. The main findings were that teachers rated most assessment characteristics as more authentic than students did, while freshman and senior students did not differ in their perception of authenticity. These findings are important for designing assessments that are authentic in the eye of the student and, in turn, promote student learning.

[4] This chapter is based on Gulikers, J. T. M., Kester, L., Kirschner, P. A., & Bastiaens, Th. J., (2006). *Authenticity is in the eye of the beholder: Student and teacher perceptions of assessment authenticity.* Manuscript submitted for publication

Student assessment has always been an important aspect of educational practice, but in the last decades a paradigm shift has taken place in assessment practices (Birenbaum, 1996; 2003; Black & William, 1998; Dochy & McDowell, 1997). This shift can mostly be seen as a reaction to societal changes (Dochy, 2001). Modern society is dynamic, knowledge is constantly changing and a lot of developments take place in the field of information and communication technologies. As a reaction to these developments, jobs have changed and different requirements are placed on employees. They are expected to be lifelong learners and be able to cope with and adapt to many different kinds of tasks and jobs. To prepare students for the world of work, education and assessment practices are also required to change.

Traditional assessment practices were referred to as belonging to a *testing culture*, while new practices belong to the so-called *assessment culture* (Birenbaum, 1996). The testing culture focused mainly on measuring of the acquisition of atomised pieces of knowledge and basic skills, while the assessment culture emphasises evaluating and stimulating the development of higher-order skills or competencies.

Segers and colleagues (2003) argue that several continua characterise traditional tests on the one hand and new assessments on the other. This chapter focuses on one of these continua, namely the authenticity continuum, with artificial and decontextualized testing on the one side and authentic and situated assessment on the other. New modes of assessment that focus on competency assessment tend to lean towards the authentic side of the continuum, since authenticity is expected to be crucial for preparing students for the dynamic world of work that characterises current society (Boud, 1990; Dochy, 2001; Segers et al., 2003).

In the sector of Vocational Education and Training (VET) in the Netherlands, learning and working in professional practice are alternated on a regular basis from the start. In this kind of education, which is gaining popularity in other countries as well, students start doing internships or apprenticeships in professional practice at the beginning of their studies. To better prepare students for this world of work, VET programmes stress the need to integrate learning and working from the start. This means that there should be a correspondence between what students have to do during learning or assessment of this learning and what students of a certain discipline are expected to do during internships or after finishing their school (Biggs, 1996; Boud, 1995; Messick, 1994; Stein, Isaacs, & Andrews, 2004; Tillema, Dekkers, & Meijers, 2000). Authentic assessments are thought to help bring learning, assessment and working closer together. Following this reasoning, this study defines authenticity in terms of its *resemblance to students' future professional practice* (Gulikers, Bastiaens, & Kirschner, 2004). By creating this resemblance, using authentic assessment during learning is thought to show students the link between learning and working in practice, thereby directing their learning towards developing professionally relevant skills (Herrington & Herrington, 1998; Lizzio & Wilson, 2004a; McDowell, 1995). However, authenticity is, at least partly, subjective , meaning that what one person perceives as authentic does not have to be perceived as being authentic by someone else. Since *student perceptions* of assessment characteristics are found to mediate the influence of an assessment on student learning (Boud, 1995; Drew, 2001; Scouller, 1995; 1997), an important question is if students (users of an assessment) and teachers (developers of the assessment)

perceive assessment authenticity in the same way. Differences or similarities with respect to assessment authenticity between students and teachers or students of different years of study might have important implications for using authentic assessments during a curriculum.

Subjective Authenticity and the Role of Perceptions

The educational goal of more authentic assessment is to stimulate deep learning activities and the development of more professionally relevant skills or competencies (Boud, 1995, Dochy, 2001, Herrington & Herrington, 1998; Tillema et al., 2000). Unfortunately, this relationship between assessment and learning is not that straightforward. Learning is influenced by assessment in three ways (Boud, 1995), namely by (a) the intrinsic or objective qualities of the assessment; (b) a teacher's interpretation of the to be assessed material. (i.e., a teacher translates the material to be assessed into a certain format and select assessment tasks appropriate for the subject and the specific learning goals), and (c) by a student's interpretation of the task at hand and the context of the assessment. Previous research (Entwistle, 1991; Gijbels, 2005; Scouller & Prosser, 1994; Scouller 1997; Struyven, Dochy, & Janssen, 2003; Van Rossum & Schenk, 1984) has shown that especially the third element, being student perceptions of the assessment characteristics, is crucial for determining learning. For example, when students perceive the assessment as measuring recall of factual information, they employ a surface study strategy that seems suitable for learning factual information by heart

With respect to assessment authenticity, student perceptions might be influential as well. The aforementioned definition that assessment authenticity depends on the degree of resemblance between the assessment and professional practice might make is seem as if authenticity is an objective construct, but authenticity is not an 'objective' quality as such. Something is only authentic with respect to something else, for example a situation, place or profession (Honebein, Duffy, & Fishman, 1993; Messick, 1994; Petraglia, 1998; Radinsky, Bouillion, Lento, & Gomes, 2001). Whether a person sees an assessment as being authentic depends on the frame of reference that person has in mind against which the authenticity is measured. In addition, a person's perception of authenticity can also change. This can be the result of the amount and kind of practical experience, schooling or age (Honebein et al; Lizzio & Wilson, 2004a; Petraglia, 1998). This means that what one person perceives as being authentic is not necessarily authentic in the eyes of someone else. In reality, thus, "authenticity is in the eye of the beholder". If it is true that student perceptions of assessment authenticity drive student learning then this indicates that before an authentic assessment can positively influence learning, it is imperative that the learner *perceives* the assessment as being authentic (Radinsky et al., 2001).

Differences in Perceptions of Authenticity

The premise behind this research is that authenticity is subjective; the perception of authenticity might be different for people of different ages or with different kinds or amounts of practical experiences. A consequence of this is that it cannot be automatically assumed that what teachers see as being authentic, and thus what they make use of in the lessons or assessments that they develop, is also perceived as being authentic by students (Radinsky et al., 2001; Roelofs &

Terwel, 1999; Stein et al., 2004). Several studies have shown that there are differences between teacher and student perceptions of a learning environment or assessment (Boud, 1995; MacLellan, 2001; Ngar-Fun, 2005; Sambell, McDowell, & Brown, 1997). They argue that teachers often use an assessment to send a message to students about what kind of learning is required, but a student's perception of this message is not always in line with the intentions of the teacher. Students create their own "hidden curriculum" (Sambell & McDowell, 1998, p. 391); they interpret the learning environment and assessment practices in their own way, which in turn drives their learning. Lizzio and Wilson (2004a), for example, showed that student perceptions of relevance of the to-be-developed skill for their future work drives their willingness and interest in acquiring that particular skill.

In light of the possible strong effect of authenticity on learning, moderated by student perceptions of this authenticity, it is important to examine how students perceive the authenticity of an assessment and if and how they differ from their teachers.

Besides the possible differences between students and teachers, students with different amounts of experience in professional practice might also differ in how they perceive assessment authenticity. In VET in the Netherlands students start doing internships from the very beginning of their studies. In other words, students are gaining much professional practice experience during their studies. Senior students, thus, have had a lot of practical experience, while freshman do not. Lizzio and Wilson (2004b) argued that students with little professional experience have unrealistic expectations about work and work roles, while seniors, having more professional experience, might have changed perceptions of work and work lives (possibly in the direction of more realistic expectations). These differences might influence what both student groups perceive as authentic assessments. This, in turn, might have important consequences for designing and using authentic assessments during a curriculum in which student gain working experience wile studying.

This Study

The main focus of this study is on examining differences in assessment authenticity perceptions of teachers versus students and freshman versus senior students. A five-dimensional framework (5DF; see Figure 3.1) for assessment authenticity (Gulikers et al., 2004) is used as a tool for describing the *objective* authenticity of the assessment used in this study and for examining assessment authenticity from the student and teacher perspectives. This framework argues that five assessment characteristics influence the degree of authenticity of the assessment as a whole. The five assessment characteristics can be described as follows:

1. Task. The assessment assignment that defines the content of the assessment
2. Physical context. The environment in which students have to perform the assessment task
3. Social context. The interaction (im)possibilities during the assessment
4. Form. The assessment method, independent of the content
5. Criteria. The characteristics of the performance (product/process) that are valued

The rationale behind this framework is that these five characteristics can resemble professional practice to a more or lesser extend. Thus, an assessment can be made more or less

authentic on five continuous scales. This framework makes it possible to describe and examine the resemblance between professional practice and these five assessment characteristics.

The research questions of this study are: (1) Do students and teachers agree about the authenticity of the assessment characteristics?, and (2) Do freshman students, with little professional practice experience, and senior students, with sizably more practical experience, differ in how authentic they perceive the assessment characteristics to be?

Method

Participants

A group of freshman students (n = 66; mean age = 18.13; SD = 1.67) and a group of senior students (n = 118; mean age = 19.16; SD = 1.14) studying Social Work at a Vocational Education and Training institute (VET) enrolled in this study. Seniors would graduate within four months, while freshman students started their studies six month earlier. Students were studying Social Work in a competency-based learning environment combined with authentic assessments. Both groups differed in their amount and kind of practical experience. Freshman students had been working in one professional setting for one day a week. Senior students had completed various internships, which spanned a continuum from one day a week to ten weeks full-time. Moreover, freshman students had no experience with assessment in the workplace, whereas senior students did. In addition, 17 teachers of the freshmen programme and 19 teachers of the senior programme participated in this study. All were involved in the authentic assessment as a developer and an assessor or role-player. There was no overlap between teacher groups.

As a precondition for examining student perceptions, the researchers explicitly selected student groups that were familiar with the kind of assessment used in the study for two reasons. First, Struyven (2005) shows that if students are unfamiliar with an assessment method, their preference for this assessment is lower than for assessment methods they are familiar with. But after having experienced the new assessment method once, their preference increased significantly. It is possible that the same kind of process holds for student perceptions of the assessment, which makes the perceptions after the second experience with the assessment a more reliable one. Second, if students are confronted with something new and unfamiliar in their learning environment, they first have to adapt to this change (Gibbs, 1992). Evaluating the learning environment might be affected by a student's ability or willingness to adapt to the changes. Evaluating an element of the learning environment that students are already used to increases the likelihood of evaluating the element of interest. Both student groups were familiar with the kind of assessment used in this study. For the freshman students it was the second time that they performed this kind of assessment, for the senior students it was the seventh time.

Materials

This study made use of two existing assessments at a VET institute for social work, both of which were designed to be authentic assessments and had the same format. The only difference between the assessments was the topic.

The assessment. The topic of the assessment of the freshman students was "dealing with conflict situations". This was one of the main competencies that students had to acquire during the course "orientation towards your own possibilities". It consisted of a case describing a situation in which the student, in the role of social worker working in a social institute, had to deal with a client who was not allowed to take part in the institute-festivities because of the family's religious background. The child was angry with his parents because he wanted to join the other children in the festivities and have fun with them while the parents and their religion forbid this. During the assessment, the students had to solve this problem during a role-play in which the teacher played the mother of the child.

The topic of the assessment of the senior students was "applying for a job" on one of three vacancies for just graduated social workers. This was one of the main competencies that students had to acquire during the course "the social work organisations: working with policy". During the assessment they had to take part in a job interview with the teacher playing the role of the employer.

The objective authenticity of the assessments was described according to the five dimensions of the 5DF and is shown in Table 5.1.

Table 5.1. The objective authenticity based on the five dimensions

Assessment dimension	Description	Value
Task	Case description of a situation that is representative of students' current internship or near professional future. Based on a core competence of Social Work as described in collaboration with the work field	++
Physical context	In school, in a classroom. An unknown teacher is the role-player. The timeframe of the assessment is ten minutes, which is not realistic since in real life the talks would stop when finished. No resources are available, which are also not likely to be used in the real life situation	-
Social context	One-on-one	
	First year: This might not resemble professional practice, since in real life at least both parents will be present and the student might involve another colleague	-
	Final year: an application is always done individually	+
Form	Role-play. Doing these kinds of talks are an important part of being a social worker making this an authentic demonstration of competence	++
Criteria	Criteria are developed in collaboration with professional practice, scored on a three-point scale, made known one week before the assessment and most are directed towards observable behaviour or talk. However, two criteria mainly focus on knowledge that students have to express	+

Two researchers independently scored the authenticity of the five dimensions based on a document analysis of the assessment material and one of the researchers observed several student performances during the assessments. Both researchers scored the five dimensions of the assessment on a 5-point scale ranging from very high degree of resemblance ('++') to almost no resemblance ('- -') between the five dimensions and professional practice. The scores of the two researchers were averaged to describe the objective authenticity of the five dimensions.

Objectively speaking, both assessments had the same authenticity for both groups, except for the social context. The individual, one-on-one context is authentic in the case of the "applying for a job" assessment, since a job interview is mostly done individually in real life as well. In the case of "dealing with conflicting situations", an individual assessment is less authentic, since in real life, a social worker might choose to deal with this problem together with a colleague and when children are involved they are likely to meet with both parents.

The instructional phase. A competency-based instructional period of nine weeks preceded both authentic assessments. During eight weeks, students worked in groups on critical professional problem situations. They had to set learning goals focusing on knowledge as well as skills/attitudes. During this training phase, students had to carry out several formative assessments. These were all role-play assignments, based on a social work related problem situation, that student had to carry out with other students. The summative assessment was based on a selection of course objectives that was translated into the assessment criteria. The assessment criteria for the summative assessment were conveyed to students one week prior to the assessment in which students were freed from obligatory educational activities.

The perception questionnaire. The renewed questionnaire for measuring perceptions of authenticity of the assessment dimensions of the 5DF was used (Gulikers et al., 2004; Gulikers, Bastiaens, & Kirschner, 2006). The 24 items of the questionnaire were contextualized in this specific assessment and examined the perceived resemblance between the five characteristics and students' future professional practice as a social worker ("The task of this assessment prepared me for my future professional life of a social worker"). Thomas and Bain (1984) showed that examining study approaches for or perceptions of a specific assessment requires a contextualized questionnaire otherwise respondents report their preferred learning approach or their perceptions of assessments in general. The items were scored on a 5-point Likert scale ranging from *totally disagree* to *totally agree*, resulting in a score for the perceived degree of resemblance between the assessment characteristics and professional practice. All scales, except for the social context scale, had a reasonable internal consistency, shown in Cronbach's alpha ranging from .70 to .82. The social context scale was left out of further analysis.

Procedure

During one week, all students took part in the assessment. Students filled in all questionnaires directly after finishing the assessment. The teacher questionnaire was almost identical to the student perception questionnaire, except that the word "I" was replaced with the words "the student".

Analysis

To examine if the groups agreed on the authenticity of the five assessment characteristics, two measurements were used. First, one-sample *t*-tests were used in which the means of the groups were individually compared to the median score of the rating scale (value "3") to find out if the four groups rated the authenticity of the assessment characteristics above or below average. Second, ANOVAs were used to examine if and how the groups differed in their ratings of the

authenticity of the characteristics. The Games-Howell post hoc test was used, since this is an appropriate test in the case of different group sizes (Field, 2000).

Results

Table 5.2 shows the mean scores on the authenticity scales for the four groups.

Table 5.2. Mean scores on the perception scales of the four groups

	Senior students (n = 118)		Freshman students (n = 66)		Senior teachers (n = 17)		Freshman teachers (n = 19)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Task	3.10	.77	3.21	.48	3.92	.67	3.86	.59
Physical context	2.53	.92	2.76	.68	2.85	1.00	2.63	.87
Form	3.31	.74	3.41	.59	4.22	.36	3.92	.50
Criteria	3.20	.62	3.26	.42	4.05	.53	3.86	.63

In line with the objective scoring of the authenticity of the assessment characteristics (see Table 5.1), all groups perceived the *physical context* as the least authentic characteristic of the assessments (Table 5.2). In addition, the one-sample *t*-tests showed the student groups rated the physical context significantly lower than three ($t(65) = 2.88$, $p < 0.01$ for first year students; $t(117) = 5.53$, $p < 0.01$ for final year students). The *task*, *form* and *criteria* of the assessment were valued as more than averagely authentic in all groups ($p < 0.01$) except for the senior students in which the rating of the task did not deviate from the median score of 3 ($t(117) = 1,47$, *ns*.). In the objective rating of the authenticity of the assessment (Table 5.1) the task, form and criteria were scored as (highly) authentic as well.

ANOVA tests showed significant differences between the four groups on the *task*, *form* and *criteria* dimension ($F(3, 216) = 12.86$, $p < 0.01$; $F(3, 216) = 12.73$, $p < 0.01$; $F(3, 216) = 17.28$, $p < .01$ respectively). Games-Howell post hoc tests showed a similar picture for the three dimensions (for Means and Standard Deviations see Table 5.2): (a) in all cases, teachers scored significantly higher than students ($p < 0.01$) (b) both student groups did not differ from one another; and (c) the two teacher groups did not differ from each other. The four groups did not differ in their perception of the authenticity of the *physical context* ($F(3, 216) = 1.38$, *ns*.).

Conclusion and Discussion

At a general level, students and teachers agreed about the question if the resemblance of the assessment characteristics and professional practice was above or below average. Their ratings at this general level were also in line with the objective ratings based on the 5DF (Table 5.1). The 5DF seems to differentiate appropriately between various characteristics that are important in developing authentic assessments. A closer look at student and teacher perceptions however showed that there are important differences that should be taken into account.

Two reasons can be used to explain the differences between student and teacher perceptions and especially the finding that teachers perceived most elements as significantly more authentic than students. First, as a result of the gap between teachers' beliefs and their actual assessment practices (Maclellan, 2001; Ngar-Fun, 2005; Orrell, 2003; Verhoeven & Verloop, 2002; Wiggins, 1989), the actual practices might be less authentic than teachers think they are, while students only see the actual practices. There is no one-on-one relationship between what teachers believe that they are implementing and what they actually implement. Second, as a result of more and different kinds of experience in professional practice, teachers are likely to have a different idea of what professional practice looks like than students do (Honebein et al., 1993; Huang, 2000; Radinsky et al., 2001; Roelofs & Terwel, 1999; Petraglia, 1998) and teachers develop assessments according to their ideas of authenticity without taking student ideas into account.

With respect to the first explanation several studies showed that there often is a gap between what teachers want to do, think they are doing and what they are actually doing (Orrell, 2003; Maclellan, 2001; Ngar-Fun, 2005; Richardson, 2005). Wigging (1989) and Maclellan even argued that the greatest disjunction between beliefs about good assessment practice and actual practice was associated with authentic assessment. Teachers often think that their assessment practices are more authentic than they actually are (Verhoeven & Verloop, 2002). Petraglia (1998) argued that a problem with authenticity is that it is not communicated explicitly. It is such an intuitive concept that people do not feel the need to be explicit about it since "everybody knows what we are talking about". Lizzio and Wilson (2004a) argued that in order to motivate student learning of professionally relevant skills, teachers should pay particular attention to student perceptions of the authenticity of the assessment and the authenticity of the skills that are required in the assessment, because students' interest in developing skills is largely dependent on their perceived relevance of the skill to their future work. Thus, teachers should explicitly discuss their intentions with a certain assessment in terms of its relevance for the future working lives of the students (Lizzio & Wilson, 2004a; Orrell). Making tacit beliefs explicit and discuss them with students possibly influences student perceptions or confronts misperceptions, and increases the likelihood that students are going to perceive the assessment as being authentic. This is important for developing better authentic assessments and for positively influencing student learning with these assessments.

With respect to the second explanation for different student and teacher perceptions, it is widely accepted that previous experiences colour the way current learning environments are perceived (e.g., Biggs, 1989; Birenbaum, 2003; Sternberg, 1999). This study hypothesised that the kind and amount of experience in professional practice influences a person's perception of authenticity. Teachers are likely to have more practical experience than students and, as a result, might perceive authentic assessments differently (Radinsky et al., 2001). A pitfall is that teachers are often the ones to develop an assessment, while students' perceptions of this assessment determine their learning. Teachers develop an assessment according to what they think is authentic for the professional field (Huang, 2002; Petraglia, 1998). Petraglia called this "pre-authentication" (p. 53) which is "the attempt to make learning materials and environments correspond to the real world prior to the learner's interaction with them". This then reflects the

real world as teachers see it, while from a learner's point of view, this might not be authentic at all, since they might have a different perception of what the real world involves (Roelofs & Terwel, 1999; Stein et al., 2004). This might be the reason that teachers rate the assessment as more authentic than the students. Cumming and Maxwell (1999), in turn, argued that when students do not perceive the intended authentic assessment as authentic, this might be detrimental for their learning. An empirical study (Cooper, 1994) supported this idea by showing that an assessment that was developed to be authentic was perceived by students as artificial and as a result inhibited their learning process.

Contrary to the expectation, freshman and senior students did not differ in their perception of the authenticity of the assessment. It was expected that the fact that senior students have much more experience in professional practice combined with experiences of assessments in the workplace, would have changed their view on authenticity in relation to the more inexperienced freshman students. Previous studies also expected to find differences between students of different years of study, but these expectations were not confirmed either (Handal & Hofgaard Lycke, 2005; Winning, Elaine, & Townsend, 2005). Winning and colleagues found that third and fifth year dental students from two institutes, who differed in their amount of practical/clinical experience and experience with different assessment methods, did not differ in what they perceived as (a) the important purposes of assessment; (b) the features of assessment important for positive outcomes; and (c) the assessment methods important for judging their learning. Handal and Hofgaard Lycke compared freshman to senior students with respect to their way of learning and to the kind of competencies that they thought were important. This study showed that also with respect to both these characteristics, students did not seem to differ. On the other hand, when the senior students were tested again after one year of working, their ways of learning and the competencies that they thought were important in professional live had changed. This might mean that student perceptions and the way of learning are relatively stable during the years of studying and more internship experience does not seem to influence this, while after finishing school, work experience seems to drastically change their ideas. Pena (1997) also argues that when students finish school and enter the professional field, they often experience a "reality shock", because they experience that the real world of work is still (completely) different from what they expected or assumed while studying. This suggests, just as Boud in 1990 already argued, that there is still a big gap between learning/assessment and working. This might explain why first and final year students do not differ, while students and teachers do differ.

Limitations and Future Research

The findings of this study suggest that students and teachers differ in their perceptions of authentic assessment, teachers perceived the most characteristics of assessment as more related to professional practice than students did, and that freshman and senior students did not differ in their perceptions of authenticity of the same kind of assessment.

The findings give food for thought for further research, but first some methodological issues need to be mentioned. First, a 5DF (Gulikers et al., 2004) was used to score the objective authenticity of the assessment, but this can never be completely objective, since it will always be

an appraisal done by a person. By using two independent raters and by doing a document analysis as well as an actual observation of the assessment, the rating was thought to be as objective as possible. Second, even though the two assessments used in this study (one for the freshman and one for the senior students) are of the same format and the objective rating of the authenticity of both assessments is almost the same (except for the social context authenticity), they are not identical, which might have influenced the findings. However, letting senior students rate the authenticity of a freshman assessment seemed inappropriate as well, as this was not expected to be very authentic for final-year students because it would be below their current educational or expertise level (i.e., the freshman assessment would be too easy for senior students and as a result not representative for their current internship or future work; Gulikers et al., 2004).

Additional research into student learning is necessary, since the relevance of differentiating between student and teachers perceptions is only meaningful if student perceptions of authenticity do indeed influence study behaviour or learning outcome. The relatively large body of evidence for the relations between perceptions of assessment characteristics and student learning seems convincing (Drew, 2001; Entwistle, 1991; Gijbels, 2005; Scouller, 1997; Scouller & Prosser, 1994; Struyven et al., 2003; Van Rossum & Schenk, 1984). With respect to authenticity perceptions, Lizzio and Wilson (2004a) showed that students' perceived relevance of a to be developed skill for professional practice is a strong predictor of their interest in developing that skill. These findings indicate a promising avenue for future research into the relationships between authenticity perceptions and student learning and development of professionally relevant skills. This, in turn, will have important consequences for designing and using assessments with certain degrees of authenticity during a curriculum. In addition, findings on student perceptions of authenticity might not only be important for assessment, but for instruction as well. In line with the constructive alignment theory of Biggs (1996), authentic assessment can only be effective in stimulating student learning when combined with authentic instruction or training. The 5DF seems to be a helpful tool for examining perceptions of authenticity and could also be used to develop assessments or instruction with different elements of authenticity.

Moreover, it would be interesting to study how student perceptions of authenticity influence student learning for first- and final-year students. It might be possible that even though they do not differ in their perception of authenticity, they might differ in the value that they place on the authenticity of an assessment (Honebein et al., 1993), which would in turn influence their study approaches or development of professional skills.

If student perceptions of authenticity do indeed influence their learning then two directions for future research and educational practice are possible. First, research should examine what makes students perceive an assessment as authentic (Boud, 1995), what influences their perception of authenticity, and find out what kind of skills they perceive as relevant for their future work (Lizzio & Wilson, 2004a). In turn, assessments that take student perceptions into account should be developed and the influence of these assessments on students learning and professional skill development should be studied. Second, instead of adapting the assessment to

student perceptions, it can also be argued that the perceptions should be influenced and changed (Van Rossum & Hamer, 2003). As said before, graduated students who enter the workforce still experience a reality shock, which suggests that students do not have a realistic perception of real working-life. It should be studied if it is possible to change student perceptions to be more in line with what is expected in the field of work. If, however, student perceptions are stable during their years of study and turn out to be difficult to change during studying, using the same effective kind of authentic assessment during a curriculum might be more efficient than putting a lot of effort, time and energy in developing all different kinds of authentic assessments during a curriculum.

Chapter 6

Relations between Student Perceptions of Assessment Authenticity, Study Approaches and Learning Outcome[5]

The study in this chapter examined the relationships between perceptions of authenticity and alignment, study approach and learning outcome. Senior students of a vocational training program me performed an authentic assessment and filled in a questionnaire about the authenticity of various assessment characteristics and the alignment between the assessment and the instruction. Deep or surface study activities and the development of generic transferable skills were measured with a questionnaire as well. Correlational analysis and structural equation modeling were used to examine the hypothesis that more perception of authenticity and alignment resulted in more deep learning and development of generic skills. Results showed that when the *task*, *physical context* and assessment *form* are more authentic and when there is more alignment there is also evidence of more deep learning and/or an increase in generic skill development. Authenticity perceptions did not affect surface learning. Contrary to expectations, more authentic assessment *criteria* resulted in a decrease in deep learning and generic skill development.

[5] This chapter is based on Gulikers J. T. M., Bastiaens Th. J., Kirschner P. A., & Kester L. (in press). Relationships between student perceptions of assessment authenticity, study approach, and learning outcome. *Studies in Educational Evaluation.*

Boud (1990, p. 101) stated that "there is often a gap between what we require of students in assessment tasks and what occurs in the world of work" and Gibbs (1992) argued that "the tail wags the dog" in that student learning is very much guided by the ways in which the learning is assessed. These two ideas show the background of this study that deals with making assessment look more like professional practice (i.e., authentic assessments) in order to stimulate students to learn and develop the knowledge, skills, and attitudes (i.e., competencies) they need for their future working lives.

An important goal of education, at least in vocational education, is to prepare students for a professional life. In the industrial era, working class people were educated for efficient functioning as skilled workers at the assembly line (Birenbaum, 2003). Schooling focused on acquiring factual knowledge and basic skills mainly through drill and practice. Current society, however, is dynamic and characterised by rapid developments in information and communication technologies and their effects on the size and sustainability of our knowledge base. Jobs have changed and different requirements are placed on graduates. Successful performance in this society demands not only a profound knowledge base and routine skills, but rather the ability to flexibly adapt knowledge and integrate it with skills and attitudes to solve new problems and handle unknown situations. To prepare students for the jobs that characterise modern society, students need to learn different "things" in a different way. As a reaction to this, the last 15 years have witnessed a lot of educational practices, at least in vocational education in the Netherlands (Tillema, Kessels, & Meijers, 2000). Schools changed their curricula and pedagogy towards more competency-based education. But changing teaching is not enough. According to the constructive alignment theory (Biggs, 1996), to change learning, both instruction and assessment practices need to change. Changing the assessments might be even more important as learning is so driven by assessment that the form and nature of assessment can swamp the effect of any other aspect of the curriculum (Boud, 1990).

Assessment Authenticity

The shift from the testing culture to the assessment culture (Birenbaum, 1996) describes several transitions in assessment practices towards assessments that fit with the ideas of competency-based education. Assessments, characteristic of the assessment culture, aim at preparing students for the dynamic world of work by increasing the correspondence between what students need to do in school and what is expected from them after finishing their studies (Boud, 1995). One of the major transitions is that these new modes of assessment are characterised by being integrated and authentic instead of atomistic and decontextualised (Segers, 2003). By brining assessments "in context" and focusing on letting students experience professional practice, authentic assessments are expected to stimulate students to develop competencies relevant for professional practice (Dall'Alba & Sandberg, 1996; Velde, 1999). Gulikers, Bastiaens, and Kirschner (2004) argued that to stimulate students to develop competencies relevant for professional practice, the assessment should require students to demonstrate the same competencies as experts would use in the real-life situation. This is more likely to occur when there is a greater correspondence between the assessment situation and the professional practice situation on which the assessment

is based. Several qualitative studies (Herrington & Herrington, 1998; McDowell, 2001; Lizzio & Wilson, 2004a) showed that students experienced an assessment to be positive for their learning when the assessment is related to authentic tasks, encourage them to apply knowledge in realistic contexts, show them relevance for their future life outside school, and/or emphasise the use and development of skills that are needed in professional life. Thus, from a theoretical as well as from a student point of view, increasing the authenticity of an assessment is expected to have a positive influence on student learning and help students prepare for their working life.

Cumming and Maxwell (2002) showed that in many educational practices the importance of authenticity is recognised, but the operationalisation of this authenticity is far from optimal. Making an assessment more authentic is mostly translated into making the assessment more realistic (e.g., having a higher fidelity) without careful consideration of what elements make the assessment more realistic or authentic. This superficial approach has resulted in assessments that were not beneficial for student learning (Cooper, 1994). Students could not appreciate the increased authenticity; instead they perceived the assessment to be more artificial, which only distracted them from an effective learning process. Despite good intentions of the developers, the assessments did not encourage students to adopt the kind of study approaches that were intended. This suggests that it is important to be careful when developing new modes of assessment, otherwise the results can be counter-productive for learning (Boud, 1990).

In order to carefully examine what makes an assessment authentic and how this influences student learning, this study builds on the literature study described in chapter 2 that unravelled the concept of authenticity. This resulted in a five-dimensional framework (5DF; see Figure 3.1 for a more elaborate description) that describes which assessment dimensions determine its authenticity. These are:

1. Task. The assessment assignment that defines the content of the assessment
2. Physical context. The environment in which students have to perform the assessment task
3. Social context. The interaction (im)possibilities during the assessment
4. Form. The assessment method, independent of the content
5. Criteria. The characteristics of the performance (product/process) that are valued.

This study examines how student perception of the authenticity of these five assessment dimensions influence student learning.

Student Perceptions and the Impact on Learning

The previous section argues for the importance of making an assessment more authentic. However, making an assessment more authentic in the eyes of the developer is not enough, since the effect of assessment on student learning seems to be mediated by *student perceptions* of the assessment requirements (Boud, 1995; Entwislte, 1991; Sambell et al., 1997; Scouller, 1997). These studies show that how students perceive the assessment, rather than the actual assessment or the teacher's intentions, affects to a large extent student learning. To be more specific, student perceptions of the assessment requirements influence their study approach (how they learn) and their learning outcomes (what they learn).

The 3P (Presage-Process-Product) model (Biggs, 1989) addresses the relationships between perceptions of the learning environment (presage variables), study approaches (process variables) and learning outcomes (product variables). Biggs argues that the influence of student perceptions can be very pervasive and that they can influence student learning in two ways. Perceptions of the learning environment can have a *direct* influence on learning outcomes, but the influence of perceptions of the learning outcome can also be *indirect* through study approach. Empirical results (Lizzio, Wilson, & Simons, 2002) supported both these relationships. They showed that positive perceptions of the learning environment had a direct positive effect on learning outcomes as well as an indirect effect on learning outcomes through stimulating a deep study approach. In addition, studies of Scouller (1997; 1998) and Sambell, McDowell, and Brown (1997) showed that students adapted their study approach when they perceived assessments as having different requirements. With respect to perceptions of assessment authenticity, McDowell (1995) and Herrington and Herrington (1998) showed that students say that an assessment positively influences their learning when they perceive it as relevant or as having a connection to reality. These results show that student perceptions are very important to consider when developing assessments. If increasing the authenticity of an assessment is thought to stimulate deep learning and help students develop professional competencies, then it is imperative that students perceive the assessment as authentic, which in turn should make students decide that a deep study approach would give the best learning outcomes.

Even though authentic assessment is expected to positively influence student learning, there has not been much (quantitative) research on the impact of *perceptions of authenticity* on student learning. This study tries to get more insight into the actual influences of perceptions of authenticity on study approach and learning outcomes. For this purpose, this study builds on the 5DF described in the previous section. By splitting up the concept of authenticity in the different dimensions described by the 5DF (see Figure 3.1), it becomes possible to gain a detailed picture of what influences student perceptions of assessment authenticity and how the perceptions of these different facets influence study approach and learning outcome. An empirical study of Gulikers, Bastiaens, and Martens (2005) that manipulated two dimensions (i.e., task and physical context) showed that the authenticity of a task and the physical context have a differential impact on student learning. This supports the idea of splitting up the concept of assessment authenticity into different facets and to examine their individual impact on student learning.

Alignment between Instruction, Learning and Assessment

Biggs' constructive alignment theory (1996) suggests that assessments should be considered as part of the learning environment. More specifically, to elicit a certain type of learning, instruction and assessment should *both* be directed towards this kind of learning (i.e., rote learning pedagogy should match rote learning assessment and competency learning pedagogy should match competency learning assessment). Empirical research by Segers, Dierick, and Dochy (2001) supported this. They showed that when students perceived a mismatch between a new kind of assessment that focused on applying knowledge to realistic problems and instruction that primarily valued memorisation, a positive effect of the assessment on study activities and

learning outcomes failed to appear. Theoretical as well as empirical evidence indicates that the effects of new modes of assessment should be examined in the light of the entire learning environment (Struyven, 2005). To this end, this study considers student perception of alignment between the instruction and the assessment next to examining the influence of student perceptions of the authenticity of the five assessment characteristics. It is expected that when students perceive a match between assessment and instruction, this will positively influence their study activities and learning outcomes, or at least will not be detrimental to them.

Research Questions

In this study, authentic assessments are defined as assessments that require students to demonstrate the same combinations of knowledge, skills and attitudes (i.e., competencies) that are applied in the professional practice situation on which the assessment is based (p. 23, this thesis). The two research questions of this study are: (1) How do student perceptions of the authenticity of an assessment influence study approach and learning outcome? More specifically, what are the direct and indirect influences of students' perceptions of the authenticity of five assessment characteristics on their study approach and their learning outcomes? (2) What is the impact of perception of alignment between assessment and instruction on study approach and learning outcome?

The quantitative, empirical study described here tries to determine whether the expected connections between perceptions of authenticity, study approach and learning outcomes do exist and if they do, how these connections work. Perception of authenticity is divided into perception of authenticity of the five assessment facets as defined by the 5DF (task, physical context, social context, form, and result/criteria) and it is hypothesised that these five perceptions affect study approach and learning outcomes individually. Furthermore, it is expected that perception of increased assessment authenticity stimulates students to use a deep approach to studying and develop generic, professional skills and that the employment of a surface study approach negatively influences the development of these skills.

With respect to the second research question it is hypothesised that when students perceive more alignment between assessment and instruction, meaning that they experience that the instruction and the assessment are aimed at the same kind of learning, they will employ more deep learning and reach a better learning outcome.

Method

Participants

One hundred and eighteen senior students (mean age = 19.16, *SD* = 1.14) studying Social Work at a Vocational Education and Training college (VET) enrolled in this study. The students were final year students and had been studying Social Work in a competency-based learning environment combined with authentic assessments for three and a half years. In other words, they were familiar with the kind of authentic assessment used in this study.

Materials

The assessment. This study made use of an existing assessment in a Vocational Education and Training institute for Social Work, which was designed to be an authentic assessment. The topic of the assessment was "applying for a job". From a teachers' point of view, this was thought to be very authentic for senior students, since they would finish school within four months, leaving them at the mercy of real professional practice. The assessment consisted of two parts: (a) writing a letter of application and a curriculum vitae for one of three social work related vacancies, and (b) taking part in a job interview based on the application letter. Both activities took place in school and the job interview was simulated in a role-play with a teacher playing the role of employer. One week before the assessment, students received a list of ten assessment criteria that focused on observable behavioral aspects. At the start of the assessment students received three descriptions of social work vacancies, one of which they could choose to be the task of their assessment. During the interview, students had to show that they could deal competently with the problem situation at hand. Students had to perform the assessment individually and their performance was observed and scored by two independent assessors on the set of criteria.

The instructional phase. A competency-based instructional period of 9 weeks preceded the authentic assessment. This period focused on the students' role as a professional. During eight weeks, students worked in groups on critical professional problem situations, for example "rights and obligations of employees", "dealing with the selection committee", or "coaching of participants". They had to set learning goals focusing on knowledge as well as skills and attitudes. During this training phase of self-study and skills training, students had to perform several formative assessments. These were all role-play assignments based on a new, but related problem case. The summative assessment (in this case "applying for a job") was based on a selection of course objectives that was translated into the assessment criteria. Although the course objectives were available from the beginning of the course, the assessment criteria were revealed one week prior to the assessment in which students were freed from obligatory educational activities.

The perception questionnaire. A renewed version of the perception questionnaire based on the 5DF (Gulikers et al., 2004; Gulikers, Bastiaens, & Kirschner, 2006) was used. The 24 items of the questionnaire all assessed the perception of the resemblance of these five assessment characteristics to (future) professional practice (e.g., "The task of this assessment prepared me for my future professional life as a social worker"). The items were scored on a 5-point Likert scale ranging from 1 (totally disagree) to 5 (totally agree), resulting in a score for the perceived degree of resemblance between the assessment characteristics and professional practice. All scales, except for the social context scale, had a reasonable internal consistency, shown in Cronbach's alpha ranging from .69 to .83. Due to its low reliability ($\alpha = 35$) the social context scale was excluded from further analysis.

Perception of alignment. The perception of alignment was measured by a 5-item questionnaire, examining whether students perceived the instruction to convey the same message as the assessment with regard to what kind of learning is valued (e.g., "During the instructional phase I

had to use my knowledge in the same way as during the assessment" or "Based on the instruction, I expected a different kind of assessment"). Cronbach's alpha for this scale was .73.

Study approach. Study approach was measured with the Revised-Study Process Questionnaire 2 Factors (R-SPQ-2F; Biggs, Kember, & Leung, 2002), a revision of the Study Process Questionnaire (Biggs, 1987). The R-SPQ-2F is a 20-item questionnaire that is more adapted to current society and modern ideas of education than the original. It was used to distinguish between two study approaches, namely a deep study approach (DSA) and a surface study approach (SSA). DSA is characterised by study activities that focus on understanding and constructing meaning of the content to be learned. SSA involves activities associated with memorisation and reproduction of atomised bits of factual information (Biggs, 1987). Several studies indicated reliable coefficients for the two scales, the items were short, all positively stated and without difficult wording. These were important considerations, since the research population involved students at the VET level and not at the higher professional or academic education level, which was the research population involved in most previous research done to validate study approach questionnaires. Moreover, the questionnaire was successfully used in previous research (e.g., Scouller 1997) to examine relationships between study approaches and learning outcomes. The original questionnaire was translated into Dutch and contextualized to the authentic assessment that was the object of this study. This contextualization was needed to examine students' study approach for a particular assessment, instead of their default or preferred study approach (Entwistle, McCune, & Hounsell, 2002; Thomas & Bain, 1984). Results indicated that the two scales of the translated version had a reasonable internal consistency in the VET context (Cronbach's alpha = .65 for SSA, and= .81 for DSA).

Qualitative learning outcome. The qualitative learning outcome was measured with a Dutch translation of the Generic Skill Development (GSD) scale of the Course Experience Questionnaire (CEQ) (Wilson, Lizzio, & Ramsden, 1997). This scale measured the extent to which students felt that a certain learning activity (in this case, studying for the authentic assessment) contributed to the development of six transferable generic skills (i.e., problem-solving, analytic skills, teamwork, confidence in tackling unfamiliar situation, ability to plan work, and written communication skills). This scale was added to the CEQ in 1997 as a reaction to the requirements of society in which students not only need to acquire content knowledge, but also need to possess skills relevant to employability and lifelong learning. Lizzio, Wilson, and Simons (2002) showed that this scale could be used as a qualitative learning outcome measure. In addition, teachers at three VET institutes confirmed that these skills were very relevant and part of the learning objectives for their students. The translated version revealed a good internal consistency in the VET context (Cronbach's alpha = .72).

Quantitative learning outcome. The quantitative learning outcome was measured by two independent assessors who, during the assessment, scored student performance during the assessment on several criteria that were placed in a scoring rubric. After the performance, both assessors discussed their scorings, which resulted in one final grade. Due to practical reasons, it was only possible to collect data on the final grade for 77 of the 118 students.

Analysis

To examine the relationships between the various variables, first correlational analyses were used. Correlations were calculated between all perception scales, deep and surface study approaches and both learning outcomes. To test the hypothesis about the influences of perceptions of authenticity and alignment on a deep study approach and the development of generic skills, Structural Equation Modeling (SEM) with AMOS was used. Contrary to regression analysis, this method is appropriate for examining direct as well as indirect effects on a dependent variable and this method can detect small changes *within* one group (Joreskog, 1993). The study examined the direct and indirect relationships between the independent variables (i.e., the perception of the authenticity dimensions and the perception of alignment), the intermediate variable DSA and the dependent variable GSD. The perception variables were not expected to influence SSA, but a negative influence of SSA on GSD was added to the model. These variables and their on theory based hypothesised relationships together make up the hypothesised model shown in Figure 6.1.

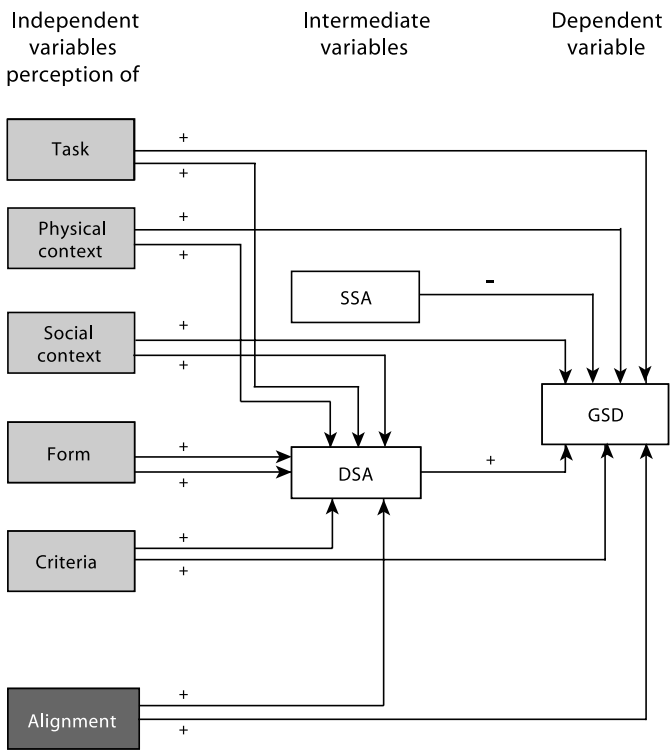


Figure 6.1. The hypothesised model showing the direct and indirect hypothesised relationships between authenticity perceptions and alignment, study approach and generic skill development (SSA = Surface Study Approach; DSA = Deep Study Approach; GSD = Generic Skill Development)

SEM was used to assess the extent to which the hypothesised model adequately fitted or described the empirical data. In this technique, several indices were used as criteria to examine the fit of the model with the data (Byrne, 2001; Joreskog, 1993). This meant that the chi-square needed to be small relative to the degrees of freedom and non-significant, the comparative fit index (CFI), normed fit index (NFI) and the goodness-of-fit index (GFI) should be large (> .95), and the root mean square error of approximation (RMSEA) should be small (< .05). To explore possible misfits of the model, the modification indexes (MI) for the regression weights could be examined (Byrne). MIs give information about the relationships that were set to zero in the tested model. High MI scores can indicate that an important link is missing in the model. The theoretical model in this study did not incorporate relations between perceptions of authenticity and a surface study approach. Therefore, these missing links were of particular interest.

For *SEM* purposes only 77 values of the dependent variable grade were available, compared to 118 valid values of the other variables. Therefore, the grade was not used in the structural model.

Results

Table 6.1 displays the correlations between the perception scales, study approaches and both learning outcomes. This table
reveals several things. First, almost all significant correlations were in the expected direction. They all stressed a positive relationship between perceptions, DSA and the learning outcomes. All perception scales correlated positively with both outcome measures, except for the physical context and grade. In addition, perception of authenticity of the *physical context* and the *task* showed a significant positive correlation with a DSA, $r(118) = .20$, $p < .05$ and $r(118) = .23$, $p < .01$ respectively. This meant that when students perceived the assessment task and/or the physical context as more authentic, they reported more use of a deep study approach. The only unexpected relationship was a positive correlation between SSA and GSD, meaning that more surface study activities improved the development of generic skills.

Second, there was a significant correlation between the GSD (qualitative learning outcome), measured with a student self-report questionnaire and the more objective grade (quantitative learning outcome), $r(77) = .25$, $p < .05$. This would imply that a higher grade coincides with more generic skill development. Third, as expected, there were no significant correlations between SSA and the perception scales, which supports the idea that authenticity perceptions do not influence surface learning. Fourth, DSA as well as SSA correlated positively with GSD, $r(118) = 52$, $p < .01$ and $r(118) = .33$, $p < .01$ respectively, while DSA correlated negatively with grade, $r(77) = .23$, $p < .05$. In other words, the employment of more deep study activities but also more surface study activities positively influenced the development of generic skills, while more deep studying resulted in a lower grade. Finally, the perception of alignment did not correlate with study approach but showed significant positive correlations with both outcome measures, $r(118) = 27$, $p < .01$ for GSD; $r(77) = .30$, $p < .05$ for grade. This would mean that perception of alignment did not influence *how* students learn (study approach), but that more perception of alignment between assessment and instruction did lead to better learning outcomes.

Table 6.1. Correlations between student perceptions, study approaches, and learning outcomes

	Surface Study Approach (SSA) (n = 118)	Deep Study Approach (DSA) (n = 118)	Generic Skill Development (GSD) (n = 118)	Grade (n = 77)
Task	-.02	.23**	.33**	.33**
Physical context	.15	.20*	.39**	.15
Form	.15	.12	.48**	.47**
Criteria	.05	-.14	.23*	.32**
Alignment	-.08	.05	.27**	.30*
Surface Study Approach	-	.02	.33**	-.10
Deep Study Approach	.02	-	.52**	-.23*
Generic Skill Development	.33**	.52**	-	.25*
Grade	-.10	-.23*	.25*	-

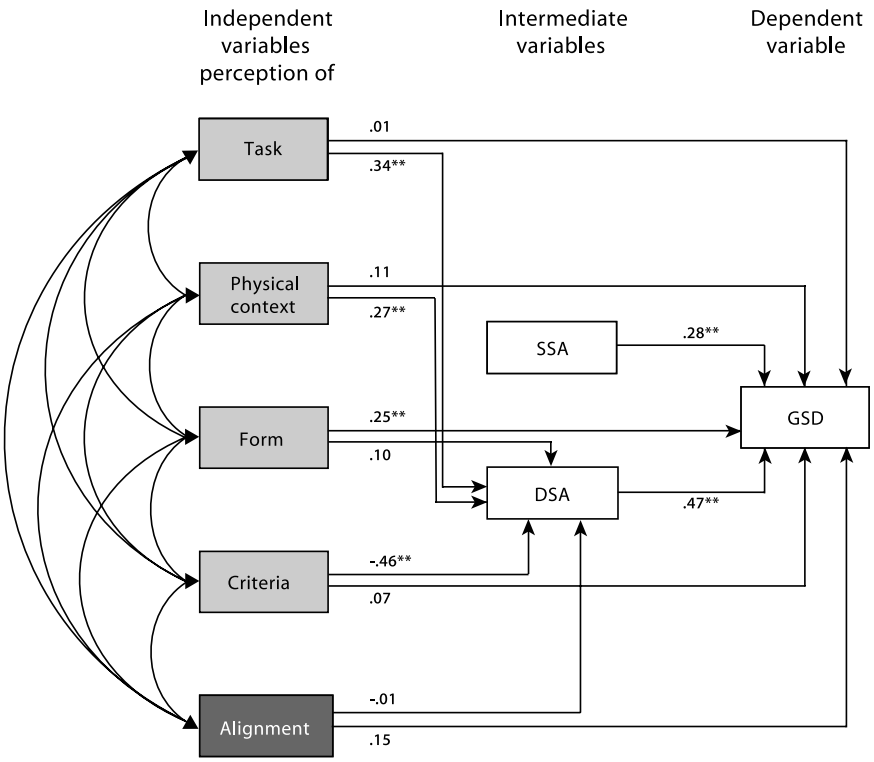
* $p < .05$, two-tailed. ** $p < .01$, two-tailed.

Figure 6.2 shows the hypothesised model and the found values of the relationships between the variables (path coefficients [β]; Note that the path coefficients are different from the correlations [*r*] between the variables displayed in Table 6.1). This model reveals several things about the influences of authenticity perceptions on study approach and learning outcome. At least four of the five authentic assessment dimensions showed significant relations with a DSA or GSD (the fifth dimension, social context, was not included due to the insufficient reliability of the scale). In line with the results of the correlations, an increase in reported DSA was seen when the assessment *task* and *physical context* were perceived as being more authentic. However, contrary to the correlations (Table 6.1), the structural model showed an unexpected significant relationship between perception of *criterion* authenticity and a DSA (β = -.46); that is, the more authentic the assessment criteria were perceived, the less deep the students reported having studied. Because the perception of criterion authenticity negatively influences a DSA, it indirectly influences the GSD in a negative way as well (β = -.22 [-.46 * .47]). The positive effect between perceived criterion authenticity and GSD that was found in the correlational analysis disappeared (r = .23, $p < .05$) when the indirect effect of perceived criterion authenticity on GSD through DSA was added to the equation.

Concerning the influences of the perception scales on GSD, almost all influences of authenticity perceptions were indirect through study approach. Only the authenticity of the assessment *form* influenced GSD in a positive and direct way (β = .25). Moreover, in agreement with the correlational analysis, both study approaches positively influenced GSD indicating that an increase in surface or deep learning both result in the development of more generic skills.

However, the influence of a DSA was almost twice as big as the influence of a SSA (β = .47 and β = .28 respectively).

The second research question dealt with the influence of perception of alignment on study approach and learning outcome. The hypothesised model shows that perception of alignment did indeed add to explaining the variance of the learning outcome (β = .15, $p < .05$), but in line with the correlations, it did not significantly influence the study approach (β = -.01, $p = .89$).



* $p < .05$. ** $p < .01$

Figure 6.2. The hypothesised model combined with the unstandardised path coefficients of the direct and indirect relationships between perceptions, study approach and generic skill development. (SSA = Surface Study Approach; DSA = Deep Study Approach; GSD = Generic Skill Development)

Conclusion and Discussion

The main hypothesis was that an increased perception of authenticity would result in more deep learning and improved learning outcomes, especially in the development of professionally relevant competencies. This was – for the most part – supported by the data. All significant correlations and the hypothesised structural model revealed positive relationships between perceptions of authenticity, a deep study approach and/or the learning outcome. More authenticity of the *task* and the *physical context* of the assessment increased the use of a deep

study approach. Increased perception of the authenticity of the *task*, *physical context*, and the *form* of the assessment all appeared to positively influence generic skill development and/or grade. In addition, an important finding was that there are no significant correlations found between perceptions and a surface study approach. This supports the adequacy of the theory-based hypothesised model used in this study that describes positive relationships between authenticity perceptions and deep learning and no relationships between perceptions and surface learning.

However, some unexpected relations were found as well. First, the structural model showed that an increase in the perception of authenticity of the assessment *criteria* negatively influenced a deep approach to studying. As a result, perceived criterion authenticity also has a negative, indirect effect on generic skill development. This is contrary to the significant positive *correlation* between perceived criterion authenticity and generic skill development (Table 6.1). This positive relationship disappeared when the indirect relationship through a deep study approach was added. This finding shows the additional value of structural equation modeling over correlations or regression that only examine direct relationships. Second, not only deep studying, but also surface studying (to a lesser extent) resulted in more generic skill development.

Two possible explanations for the negative relationship between authentic criteria and deep studying might be (1) that the criteria were too specific and/or (2) that they were revealed only one week before the assessment. The criteria focused on very specific and concrete behavioural actions such as, for example, "the student makes eye contact with the client" or "the student asks at least one open question". Previous research (e.g., Govaerts, Van der Vleuten, Schuwirth, & Muijtjens, 2005) showed that too specific criteria that focus on small concrete actions were demotivating for students further in their educational trajectory. This demotivation might in turn inhibit learning. Second, students did not receive the assessment criteria at the beginning of the course, but only one week prior to the assessment. In other words, students did not get the criteria when studying, but rather when they had one week off to focus only on preparing for the assessment. This might stimulate students to focus especially on these selected criteria as Boud (1990) argued that assessment encourages students to focus on the topics that are assessed at the expense of those that are not. The concreteness or specificity of the assessment criteria and the fact that students received them only one week before the assessment possibly stimulated learning the criteria by heart and practicing in demonstrating only these specific actions, instead of learning in a more holistic way focused on understanding (deep study approach). The assessment culture advocates transparent and concrete criteria (e.g., Dierick & Dochy, 2001) to let students know what is expected of them, but this study shows that this can have a negative effect on student learning if not implemented or perceived correctly. Moreover, one can question if performance criteria in real life are always that specific and concrete (Hager, Gonczi, & Anthanasou, 1994).

The finding that deep learning as well as surface learning positively influenced generic skill development shows that students employing a deep study approach were able to effectively deal with the assessment, but that students who mainly used surface activities could get by as well. Thus, succeeding in this assessment did not require deep learning. This result was found in previous studies as well (Biggs, 1987; Gijbels, 2005; Scouller & Prosser, 1994). These studies

showed that although a deep study approach is expected to lead to higher achievement (both in terms of quality as well as quantity) and new kinds of assessments are expected to require a deep study approach, assessment does not always reward the deep approach. On the other hand, the positive influence of deep learning on the learning outcome is stronger than the influence of surface learning, which seems to indicate an advantage for deep learners. Another explanation for the positive effect of both surface and deep learning on generic skill development is given by the four-component instructional design model for developing complex skills (Van Merriënboer, 1997). This model states that acquiring complex skills requires deep understanding of the non-routine aspects of a complex skill as well as memorisation and drill and practice of the routine aspects of a complex skill. In other words, surface study activities as well as deep study activities are required in complex skill development.

The second research question considered the positive influence of perception of alignment on learning. The study showed that if students perceive that both instruction and assessment focus on the same kind of learning, this does not influence their study approach but it does positively influence their learning outcomes. This seems to hold for any kind of learning, since the alignment scale measured if students thought that the instruction and the assessment required the same kind of learning, without referring to a specific kind of learning. This corroborates the theory and empirical data on the need for constructive alignment (Biggs, 1996; Segers et al., 2001) and suggests that it is always valuable to examine the effect of assessment on students in the light of the whole learning environment of which the assessment is part.

The significant correlations and path coefficients seem small indicating a small to moderate effect size (correlations ranging from $r = .20$ to $r = .52$ and path coefficients ranging from $\beta = .15$ to $\beta = .47$). Cohen (1988) argues that we should compare the values to other, comparable studies in the field, since the found effects in behavioural sciences will always be much smaller compared to the effects found in, for example, the physical sciences. If we look at the data from this point of view, the found effects are not that small at all. Lizzio, Wilson, and Simons (2002) examined direct and indirect relationships between perceptions of the learning environment (including perceptions of the assessment), deep and surface study approaches, and learning outcomes (grade and generic skill development). They reported path coefficients ranging from $\beta = .07$ to $\beta = .32$. Tang (1991) described relations between general study approach, assessment preparations strategies and learning outcomes and found coefficients ranging from $\beta = .10$ to $\beta = .43$.

An additional, but important finding in this study was the significant positive correlation between the qualitative and quantitative learning outcomes (general skill development and grade). Lizzio and colleagues (2002) argued that the general skill development scale was a valid indicator of learning outcome. However, this is a self-report questionnaire, which is not always considered to be a reliable indicator for actual behaviour. The grade, on the other hand, was based on a rating of student performance by two independent assessors. The positive and significant correlation between the grade, based on assessor evaluation, and the self-reported development of generic skills corroborates the validity of the generic skill development scale as a measure of qualitative learning outcome. This might be a valuable finding, because evaluating

students qualitatively and on their development of generic skills relevant for employment will become even more important in competency-based education.

Limits and Future Directions

Some notes of caution should be drawn here. The results of the structural equation modeling should be treated with caution. Structural equation modeling was used to get a deeper insight into the (values of the) direct and indirect relationships between perceptions, study approach and learning outcomes than is possible with correlational analysis or regression. However, the model was tested with a group of 118 students who all worked with the same authentic assessment. The smaller the group of participants, the more the structural equation modeling results are dependable on the specific dataset (Byrne, 2001; Joreskog, 1993). This implies that the relationships and values found in the hypothesised model are indicative and only applicable to this student group. Future research should replicate this kind of study to examine the stability of the relationships found in this study in other cases (other students and other assessments).

Since the data set for grade was 77, structural equation modeling was only used to test the influences of perceptions on study approach and generic skill development. The correlation matrix (see Table 6.1) showed that the pattern of significant correlations between perceptions and both learning outcomes look alike, while the relationships between the study approached and both learning outcomes are different. Previous research also showed that the influences of perceptions or study approaches on a qualitative or quantitative learning outcome differed (e.g., Gijbels, 2005; Lizzio et al., 2002; Scouller & Prosser, 1994). Future research should examine the relationships between authenticity perceptions, study approach and grades. Especially as long as grades stay one of the most often used measurement of learning.

To gain a deeper insight into the relationships between student perceptions of authentic assessment and the way students study for this assessment and what they learn from it, qualitative research should be used in addition to quantitative studies. Previous research showed that asking students how they studied for a certain (kind of) assessment revealed a lot of useful and valuable information (e.g., Sambell et al., 1997), which was necessary to build a rich, and contextualized picture of the assessment under investigation. Semi-structured interviews with groups of students can reveal several explanations for the (un)expected relationships found in this study (Morgan, 1997).

Practical Value

The hypothesised structural model corroborates the premise that different facets of authenticity influence study approach and learning outcome differently. This supports the 5DF which argues that an assessment can be made more authentic in different ways and that there is not an 'authentic - not-authentic' dichotomy, but rather an authenticity continuum. In agreement with Gibbs (1999), the 5DF and this study show that making small changes to the assessment (e.g., increasing the authenticity of the task) can positively influence student learning. In practical terms, this research shows that changing from traditional to competency-based education with authentic assessments need not to be a "one-shot deal". This would make the transition a lot easier for schools and their teachers. For example, much could be gained by first increasing the

authenticity of the task, since increasing task-authenticity stimulated deep learning and resulted in better learning outcomes. In the end, however, the results of this study would argue that increasing the authenticity of the task, the physical context, *and* the assessment form results in the most benefits in terms of learning and outcomes.

The study does support the idea that authentic assessment is a multidimensional concept and that various aspects of authenticity influence what and how students learn in a competency-based curriculum in which assessment, instruction and learning are in alignment and aim at bridging the gap between learning and working.

Chapter 7

The Influence of Practical Experience on Perceptions, Study Approach and Learning Outcomes in Authentic Assessment[6]

Does authentic assessment, or the perception thereof, affect how students study and learn? Does practical experience affect how assessment authenticity is perceived? And does practical experience influence how an authentic assessment affects student learning? Mixed methods design yielded insight into the answers to these questions. The study in this chapter examines the authenticity perceptions as well as the relationships between authenticity perceptions of different cohorts of students, who differ in the amount of practical experience, their study approach and ultimately their degree of professional skill development. The results show, generally speaking, that perceptions of authenticity and the influence of these perceptions on student learning are fairly stable during the years of study, with a few, very salient differences. These results suggest guidelines for developing and using authentic assessments during a curriculum in which learning and working are intertwined.

6 This chapter is based on Gulikers J. T. M., Kester, L., Kirschner P. A., & Bastiaens Th. J. (2006). *The influence of practical experience on perceptions, study approach and learning outcomes in authentic assessment*. Manuscript submitted for publication.

One of the main characteristics of new modes of assessment that focus on higher-order skills or competencies that are relevant for successful job performance, is that they are authentic (Boud, 1990; 1995; Dierick & Dochy, 2001; Gielen, Dochy & Janssen, 2003; Messick, 1994; Segers, 2004; Tillema, Kessels, & Meijers, 2000). Such authentic assessment aims at linking learning and working by creating a correspondence between what is assessed in the school and what students need to do in the workplace (during an internship or after finishing their education) (Boud, 1995; Cummings & Maxwell, 1999; Gulikers, Bastiaens, & Kirschner, 2004; Kasworm & Marienau, 1997; Messick). By creating this correspondence, authentic assessments are expected to positively influence student learning and better prepare students for their future careers. Authentic assessments are expected to (a) stimulate students to learn more deeply (Birenbaum, 1996; Dochy & McDowell, 1997; McDowell, 1995; Frederiksen, 1984); (b) stimulate students to develop professionally relevant skills and thinking processes used by professionals (Gielen et al; Savery & Duffy, 1995); and (c) motivate students to learn by showing the immediate relevance of that what is learnt for professional practice (Herrington & Herrington, 1998; Lizzio & Wilson, 2004a; Martens, Gulikers, & Bastiaens, 2004; McDowell).

The Relationships between Perceptions and Student Learning

The influence of authentic assessment on student learning, however, is not this straightforward for two reasons. First, authenticity is not an objective construct (Honebein, Duffy, & Fishman, 1993; Petraglia, 1998). This means that people can differ in their perception of the authenticity of the same assessment. The problem in this case is that *student perception* of assessment characteristics seem to have more influence of student learning than the *objective* characteristics themselves (Entwistle, 1991; Struyven, Dochy, & Janssen, 2003; Van Rossum & Schenk, 1984). This implies that if students can perceive the authenticity of an assessment differently, then the influence of this assessment on student learning can be different as well. If assessment authenticity is defined by the degree of correspondence between the assessment and the professional practice situation it means to reflect (Gulikers et al., 2004), then the influence of an authentic assessment on student learning depends on how a student perceives the resemblance between this assessment and professional practice.

Second, a student's ideas of professional practice might change as a result of more or different experiences in professional practice (Lizzio & Wilson, 2004a; Handel & Hofgaard-Lycke, 2005; Honebein et al., 1993). Following the previous line of reasoning, changed ideas about professional practice might change the perception of authenticity of an assessment, which, in turn, might influence how an authentic assessment influences learning by students with different amounts of experience in professional practice. Whether or not this is true is a highly important question in current educational practice. Many types of education, especially vocational, are trying to integrate learning and working in order to smoothen the transition from school to the workplace (Boshuizen, Bromme, & Gruner, 2004; Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004). This approach entails increasing the student's opportunity to gain experience in professional practice during schooling. The opinions on and evidence for the effect of this experience in professional practice, however, are ambiguous.

In support of the influence of practical experience, Honebein and colleagues (1993) and Messick (1994) suggested that students with different levels of practical expertise might learn better with different kinds of assessments. As an example, they argued that when students have had enough opportunity to get a good picture of professional practice, the physical context of an assessment might become self-imposed. In other words, experienced students would be able to create a realistic physical context for themselves and do not need assessments to be situated in a high fidelity context to stimulate their learning. Inexperienced students might not yet be able to frame an assessment task in a realistic context, because they have had too little practical experience for this framing. This implies that inexperienced students benefit more from a contextualized assessment than more experienced students. On the other hand, it has also been suggested that as students get closer to their graduation and thus closer to actually working, the need for a very authentic physical context (preferably the real workplace) increases, instead of decreases, and is needed to positively influence learning (Klarus, 2000). One way or the other, all of these studies imply that the same authentic assessment differentially influences the learning of students with varying amounts of practical experiences. Specifically, the same authentic assessment might be effective for students with little experience in professional practice, while being less effective for students who have more professional experience or vice versa. In practice, this would favour the use of different kinds of authentic assessments for students with different amounts of experience in professional practice.

Other research (Handel & Hofgaard-Lycke, 2005; Winning, Elaine, & Townsend, 2005) suggested that there is no need to change the kind of authentic assessment when students gain more experience in professional practice, because students do not change their perceptions of professional practice and of assessments during their years of studying, even if they gain more experience in professional practice through internships. Handel and Hofgaard-Lycke showed that students' perception of professional practice as well as their approach to studying did not change during the years of study, but that they did change drastically after one year of work. It was argued that as long as students stay in school they are mostly guided by school requirements rather than by their possibly changed ideas of professional practice or assessment. Important to note is that both these studies were conducted at the university level. The internship regimes are likely to be different from the internship regime in vocational education in the Netherlands, the context of this study, in which students start doing internships from the start of their studies and where learning and working are alternated on a regular basis. In addition, academic education is not vocationally oriented. However, based on the results of these studies, it is expected that the relationships between perception of assessment authenticity and student learning and professional skill development are stable over the years. In other words, the same authentic assessment would be expected to have the same influence on student learning and professional skill development.

This study focused on examining perceptions of assessment authenticity and the influence thereof on student learning. A growing body of literature and research on new modes of assessment stresses that the effects of assessments on student learning should always be examined in the light of the whole learning environment along with student perception of the

learning environment (Biggs, 1996; Birenbaum, 1996; Segers, Dierick, & Dochy, 2001; Struyven, 2005). This approach stresses the need for alignment between instruction and assessment to positively influence student learning. This means that when authentic assessment is directed towards stimulating deep study activities and the development of professional skills, then instruction should also be perceived to require these same things (Gulikers et al., 2004). When students perceive a mismatch between the kind of learning stimulated by the instruction and the kind of learning that is needed for the assessments, the expected positive effect on learning does not occur (Gijbels, 2005; Segers et al., 2001; Struyven, 2005). Therefore, this study not only considers student perceptions of assessment authenticity and the influence thereof on learning, but also their perceptions of alignment between the authentic assessment and instruction.

The Hypothesised Model

This study build on the hypothesised model that is tested in the study reported on in chapter 6 (see Figure 6.1). However, instead of testing this model *within* one student group, the current study focused on testing the stability of the relationships within this model *between* two student groups.

The independent variables are *perception of assessment authenticity* and *perception of alignment* between instruction and assessment. These are depicted in the left column of Figure 6.1. Assessment authenticity is defined as being multidimensional based on five assessment dimensions, namely the assessment task, the physical context of the assessment, the social context of the assessment, the assessment form and the assessment criteria (five-dimensional framework of assessment authenticity; 5DF, see Figure 3.1). The perception of authenticity should also be rated along these five dimensions and they are therefore depicted as five separate variables in the model.

The dependent variable is the *generic skill development* (GSD), which indicates to what extent students feel that an assessment stimulates the development of six generic professional skills, namely problem-solving, planning, collaborating, communicating, dealing with unknown situations, and thinking analytically (Wilson et al., 1997; Lizzio et al., 2002).

The intermediate variable *study approach* is split up in deep study approach (DSA) and surface study approach (SSA) (Biggs, Kember, & Leung, 2002), both depicted in the middle column and thus, situated between the independent perception variables and the dependent variable GSD. In line with the goals of authentic assessment, this model hypothesised that an increase in the perception of authenticity of any of the five assessment dimensions and/or of alignment, results in increased development of generic skills either directly or indirectly through encouraging a deep study approach.

This Study: The Influence of Amount of Practical Experience

This study first examines whether two student groups that differ in the amount of practical experience during their education perceive the authenticity of the same kind of authentic assessment equally, if they employ the same degree of surface and deep study activities and if they develop the same degree of generic skills in response to the assessment. Second, this study

examines if the hypothesised model adequately describes the *relationships* between perceptions, study approach and generic skill development and if these relationships are stable or variable by comparing the relationships found in both student groups. If differences are found, the causes of this variability are examined and it is studied whether the differences can be explained by the fact that both student groups differ in the amount of practical experience.

The research questions are thus:

1. Are the perceptions of authenticity and alignment, study approaches, and development of generic skills the same or different for freshman students, with little experience in professional practice, and senior students, with more professional practice experience?
2. Does the hypothesised model adequately describe the relationships between perceptions of assessment authenticity and alignment, study approach and the development of generic skills for both student groups?
3. Are the expected relationships stable across the two student groups that differ in the amount of practical experience?
4. If variabilities are found, then what causes the differences between groups?

Whether student perceptions and the influence thereof on learning are stable or variable throughout an educational career, has implications for the way authentic assessments should be implemented in a curriculum in order to stimulate student learning and development of professional skills during a student's educational career.

Method

Participants

Two groups of students from a Vocational Education and Training (VET) college for Social Work in the Netherlands participated in this study; 81 freshman students (mean age = 17.82, *SD* = 1.57) and 118 senior students (mean Age = 19.16, *SD* = 1.14). These two groups were selected because they reflected the moderator variable "amount of practical experience". Freshman were at the beginning of their studies and had little experience in professional practice. They had experience with only one institute where they have been doing an internship one day a week for half a year. Senior students were almost at the end of their studies and had a lot more and different kinds of practical working experience. They did internships in several institutes and of varying durations (varying from one day a week through six months full-time).

Materials and Data Collection

The assessment. Both student groups performed the same kind of authentic assessment, namely a 10-minute role-play, situated in a classroom, based on a social work related case description that students could prepare beforehand. A teacher played the role of client and students were handed a list of ten assessment criteria one week before the assessment. Students individually had to show their competence in dealing with the problem situation described in the case description. During their performance, students were observed and assessed by two assessors.

The perception questionnaire. A 24 item perception questionnaire based on the 5DF for assessment authenticity was used to examine to what degree students perceived the five assessment characteristics (the task, the physical context, the social context, the form, and the criteria) to resemble professional practice. The items were scored on a 5-point Likert scale ranging from 1 (*totally disagree*) to 5 (*totally agree*). All scales, except for the social context scale, had a reasonable internal consistency, shown in Cronbach's alpha ranging from .63 to .83. The social context scale was excluded from further analysis due to its low reliability (α = .48)

Perception of alignment. The perception of alignment was measured by a 5-item questionnaire, examining whether students perceived the instruction to convey the same message as the assessment with regard to what kind of learning is valued (e.g., "During the instructional phase I had to use my knowledge in the same way as during the assessment" or "Based on the instruction, I expected a different kind of assessment"). Cronbach's alpha for this scale was .75.

Study approach. Study approach was measured with the Revised-Study Process Questionnaire - 2 Factors (R-SPQ-2F; Biggs, Kember, & Leung, 2002), a revision of the Study Process Questionnaire (Biggs, 1987). The R-SPQ-2F is a 20-item questionnaire that is more adapted to current society and modern ideas of education than the original. It was used to distinguish between two study approaches, namely a deep study approach (DSA) and a surface study approach (SSA). The original questionnaire was translated into Dutch and contextualized to the authentic assessment that was the object of this study. Results indicated that the two scales of the translated version had a reasonable internal consistency in the VET context (Cronbach's alpha = .66 for SSA, and= .81 for DSA).

Qualitative learning outcome. The qualitative learning outcome was measured with a Dutch translation of the Generic Skill Development (GSD) scale of the Course Experience Questionnaire (CEQ) (Wilson, Lizzio, & Ramsden, 1997). This scale measured the extent to which students felt that a certain learning activity (in this case, studying for the authentic assessment) contributed to the development of six transferable generic skills (i.e., problem-solving, analytic skills, teamwork, confidence in tackling unfamiliar situation, ability to plan work, and written communication skills). Lizzio and colleagues (2002) showed that this scale could be used as a measure for qualitative learning outcome. The translated version revealed a good internal consistency in the VET context (Cronbach's alpha = .74).

Focus Groups. To gain a deeper insight into the perceptions of authenticity of the assessment, the way these perceptions influenced study approaches and learning outcome, and the influence of amount of practical experience on these relationships, the quantitative data were complemented with qualitative data obtained from semi-structured focus-group interviews with freshman and senior students. A random selection of students participated in five interviews; two freshman groups (total $n = 18$) and three senior groups (total $n = 27$). In these group interviews, participants were encouraged to freely express their perceptions and experiences and to respond to each other, based on initial stimuli provided by the interviewer. The interview schedule used in this study was based on a combination of the five-dimensional framework for assessment authenticity (Gulikers et al., 2004), and the interview schedules of Sambell, McDowell, and Brown (1997) and Lizzio, Wilson, and Simons (2002) focusing on perceptions of

assessment characteristics and on the consequential validity of the assessment. The interviews focused on (1) how students perceived the authenticity of the five characteristics of the assessment; (2) how they prepared for the assessment and if this depended on the authenticity of the assessment characteristics; and (3) what kind of learning they believed was being assessed in this way. The focus-group method is expected to generate richer information and get more in-depth insight than would be possible with questionnaires or individual interviews (Morgan, 1997; Sambell et al.).

Quantitative Analysis

T-test. T-tests were used to compare the mean levels of perception of authenticity of the assessment characteristic and alignment, study approaches, and generic skill development between the two student groups.

Structural equation modeling. To examine if more perception of authenticity and alignment resulted in more deep learning and generic skill development, structural equation modeling with AMOS (Byrne, 2001) was used. This analysis was chosen because - as opposed to regression analysis - it is appropriate for examining direct as well as indirect relationships between *continuous variables* and detecting small changes *within* one group (Joreskog, 1993). It was tested if the hypothesised model (Figure 6.1) adequately described the relationships in both students groups. To examine if the relationships in this model are equal or different between freshman and senior students, multi-group structural equation modeling was used. This method first examined the relationships *within* each group separately and then compared the found relationships *between* groups. The goal was to make inferences about group (freshman vs. senior) *differences* in the relationships. This is done in a three-step manner (Byrne):

1. Assess the tenability of the hypothesised model for each group separately.
2. Assess the tenability of the hypothesised model simultaneously across both groups (baseline model).
3. Assess group differences among individual parameters linking the various variables:
    - Constrain all theoretically interesting parameters to be equal across groups and compare this to the baseline model.
    - Sequentially release constraints if the Modification index indicates a significant improvement in data-model fit. Parameters whose constraints were released, were inferred to differ across groups; those whose constraints were not released, were inferred to be stable across groups.

To test the fit of the model (i.e., if the model is an accurate description of the data), a combination of several fit indices was used (Byrne, 2001). The chi-square needed to be small relative to the degrees of freedom and non-significant, the comparative fit index (CFI), normed fit index (NFI) and the goodness-of-fit index (GFI) should be large (> .95), and the root mean square error of approximation (RMSEA) should be small (< .05).

Qualitative Analysis

The focus-group data were used to complement the quantitative data, but they were especially used to explain differences between groups or to explain unexpected findings. The interview

data for analysis were first parsed in fragments and coded based on themes of the interview schedule used. To minimise the influence of personal interpretation and increase the reliability of the conclusions, first a selection of the interviews was coded and interpreted per theme by two researchers independently to find a common ground for coding and interpreting the data. After that, the interpretation of the remaining interviews rested upon careful reflection and discussion between two researchers, one of whom was not involved in conducting the interviews.

Results

Quantitative Results

T-test. Table 7.1 displays the means of both groups.

Table 7.1. Perception of authenticity and alignment, study approach and generic skill development of senior and freshman students

	Senior students (n = 118)		Freshman students (n = 81)	
	M	*SD*	*M*	*SD*
Task	3.10	.77	3.33	.61
Physical context	2.53	.92	2.99	.86
Form	3.31	74	3.41	.67
Criteria	3.20	.62	3.23	.46
Alignment	3.41	.74	3.28	.63
Surface study approach	2.64	.51	2.77	.50
Deep study approach	2.85	.64	2.93	.57
Generic skill development	2.81	.61	3.10	.59

The *t*-tests showed that both student groups did not differ in their perception of authenticity of the assessment form, the assessment criteria, and alignment between assessment and instruction. Also the degree to which deep or surface study approaches were reported, was stable. On the other hand, freshman students perceived the assessment task and physical context as more authentic than senior students ($t(197) = 2.22$, $p = .03$; $t(197) = 3.51$, $p = .00$ respectively), and they differed in amount of generic skill development ($t(197) = 3.18$, $p = .002$) in favour of the freshman students. Thus, freshman students perceived the assessment task and physical context as more authentic than senior students and in the end reported more generic skill development, while both groups did not differ in their study approaches.

Structural equation modeling. First, the hypothesised model was tested as a whole and after that, all the parameters were compared individually. Table 7.2 displays the data of the steps taken in the multi-groups *SEM* analyses. The hypothesised theoretical model turned out to be tenable for both groups separately ($\chi^2(6, n = 118) = 7{,}71$, $p = .26$, CFI = .99 and RMSEA = .05 for senior students and ($\chi^2(6, n = 81) = 1.52$, $p = .96$, CFI = .99, and RMSEA = .00 for freshman students). In addition, the hypothesised model was tenable for the groups together

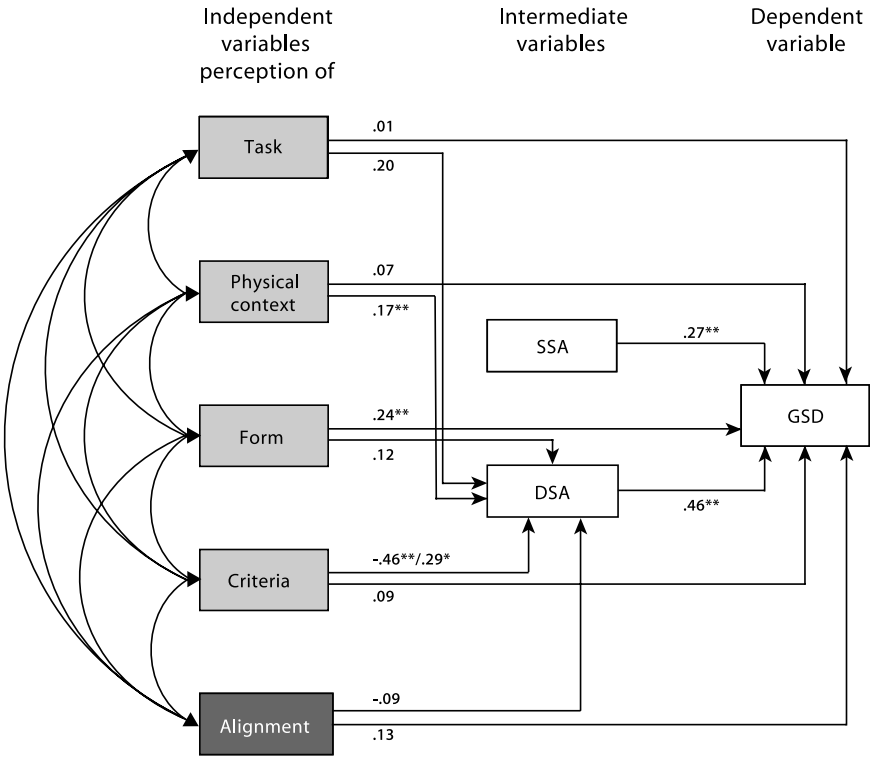
($\chi^2(12, N = 199) = 9.22$, $p = .68$, CFI = 1.0 and RMSEA = .00). This means that, within both groups, students who perceived the assessment as a whole as more authentic reported employing more deep learning and developing more generic skills compared to students who perceived the assessment as less authentic.

Table 7.2 Goodness-of-fit statistics for tests of invariance across freshman and senior students: A summary

Model description	Comparative model	χ^2	*Df*	$\Delta\chi^2$	Δdf	p
Theoretical model for 118 senior students		7.71	6			
Theoretical model for 81 freshman students		1.52	6			
Combined baseline model 118+81 (MODEL 1)		9.22	12			
All relations constrained equal	Model 1	51.04	33	41.82	21	< .05
Factor loadings constrained equal (MODEL 2)	Model 1	42.50	24	33.50	12	< .05
Factor loading constrained equal except Criteria→ DSA	Model 2	20.63	23	21.81	1	< .05

To test the equality of the individual parameters linking the independent, intermediate and dependent variables in the hypothesised model for both student groups, the values of these parameters were constrained to be equal across groups. The χ^2 of this constrained model ("model 2") was then compared to the hypothesised model ("model 1"). The χ^2 difference turned out to be significant ($\Delta\chi^2(12) = 33.50$, $p < .05$), which meant that not all parameter values were equal across both student groups. To locate the non-equivalent parameter in the model, the Modification Indexes (MI's) were studied. The MI's showed that the link between criterion authenticity and DSA did not have an equal value for the freshman and senior student groups. Releasing the equality constraint of this parameter, thereby allowing the link between criterion authenticity and DSA to differ between groups, resulted in a significant improvement of the model ($\Delta\chi^2(1) = 21.87$, $p < .05$). This meant that the value of the influence of perception of criterion authenticity on DSA, and indirectly on GSD, differed for freshman and senior students. For freshman students the value of the relationship was $\beta = .29$, while for senior students this was $\beta = -.46$. This showed that, in line with the expectations, an increase in perception of criterion authenticity stimulated freshman students to more deep studying and development of generic skills. Senior students, on the other hand reported employing less deep studying and developing generic skills when they perceived the assessment criteria as more authentic. Noteworthy was

that in both groups the perception of criterion authenticity had the strongest influence on deep learning (see Figure 7.1), but the influence was negative for seniors and positive for freshman students. There were no other parameters that differed between groups (no other modification indexes were significant). Figure 7.1 displays the results of the multi-groups *SEM* analyses.



* $p < .05$. ** $p < .01$

Figure 7.1. The structural model with the unstandardised path coefficients (DSA = deep study approach; SSA = surface study approach; GSD = generic skill development)

In short, students did not differ in perception of authenticity of the assessment form and criteria or degree of deep and surface learning in response to the same kind of assessment. However, they did differ in the average perception of authenticity of the task and physical context of the assessment and their average degree of generic skill development in response to this assessment. Also, the expected relationships as displayed in the hypothesised model were for the most part supported for both student groups, suggesting that the influence of the perception of authenticity on study approach and learning outcome is fairly stable. Students in both groups who perceived the authenticity of the task, physical context or form as more authentic, displayed more deep studying and development of generic skills than students who perceived these assessment characteristics as less authentic. Finally, the influence of perception of criterion authenticity on deep studying was strongest for both groups, but in opposite directions. For freshman students, an increased perception of criterion authenticity resulted in

more deep studying and development of generic skills, while, contrary to the expectations, this relationship was reversed for senior students.

Qualitative Results

The focus-group data were used to explain the differences found between both student groups and the unexpected findings in the quantitative analysis. Illustrative quotation from both freshman and senior students were given to support agreement or dissimilarity between the student groups. Fragment number combined with a "f" for freshman and "s" for senior focus group and the interview number are given in parentheses.

Perceptions of Task and Physical Context Authenticity. The authenticity of the assessment task referred to the degree to which the content of the assessment resembled activities of professional practice. All student groups agreed that the task of the assessment referred to activities and problems that students encounter in their internships "the cases are realistic, they are real. Look, we are social workers and these cases are really directed at the activities of social workers (73, s4)" or "[are the cases realistic?] yes definitely, you can encounter them anytime at your workplace (56, f2)". A difference that was found was that the senior groups favoured assessment tasks that they could personalise to fit with their own interests or working context or assessment tasks that dealt with extreme, more specialised cases instead of the general social work activities reflected in the current assessment. Seniors said that they would learn more from these kind of assessment tasks and be more motivated to learn for them: "I think it would be better if you can choose your own tasks … I think you learn more from situations that you really have problems with (59, s5)" or:

> [in the case of dealing with unwanted behaviour] if the case would deal with real, hard aggressive situations, then I would learn more from it or be more willing to study for it. For example, that you have to deal with a really aggressive client who is throwing chairs, instead of someone who is writing in his agenda while you are speaking, which is the kind of unwanted behaviour you get in the assessment (118, s3).

With respect to the physical context, both freshman and senior students agreed that they would respond differently in real professional practice than in a role-play. However, freshman felt that role-play assessments were appropriate for assessing job performance: "It assesses how you would respond in a certain situation, when it would take place in practice (8, f1)" or "instead of writing things down on a piece of paper, a role-play resembles a real situation and asks you to bring it [theory] in practice (61, f1)". Seniors, on the other hand, felt that assessing with role-plays in school is redundant: "I think that they should not do such an assessment, they just need to come to our workplace (88, s4)", "We already have a lot of work and life experience that you think: 'do we really have to do this in school?' For us it often feels as if we are making up things [role-plays] just to stay busy (108, s3)".

In both groups, the criteria had the greatest influence on student learning (see Figure 7.1), but the relationships were opposite for both student groups. When asked "what determines your study activities for the assessment and your behaviour during the assessment?" all student groups agreed that this were the criteria: "You just have to perform all the steps that are

described in the criteria. In this case, about a Jehovah's witness, that meant that you actually did not have to know anything about what a Jehovah's witness is (51, f2)", or "You get that list with criteria and you are completely focused on performing these points, otherwise you will fail (15, s3)". Also, both groups agreed that the criteria were realistic, in essence and on a broad level: "I think they are good points. At my workplace they also pay attention to these points (41, f1)", but that the concretisation of the criteria was too specific, theoretical, or based on schoolbooks or rules: "Theoretically speaking the criteria are really good, but it does not always work like that in practice (16, s5)", "in the assessment, I have to meet all kinds of school rules like explicitly appointing behaviour-feelings-consequences, while in practice I will not always explicitly mention how I feel or what the consequences are (61, f2)", or "you use them [the criteria] in practice all the time, but still it is different from how you've learnt it here in school. You use them, but you just translate them in their [your clients] language (23, s5)".

Differences were found in how the realistic but specific criteria influenced freshman or senior student learning. In short, it seemed as if freshman felt more comfortable with specific criteria, because this gave guidance and structure to their learning, whereas seniors felt that the criteria were appropriate on a general level, but the specification inhibited them to respond naturally. Freshmen students said: "I think this description of the criteria is logical, otherwise you don't know… I mean, they are just 'helping points' like: there are three causes mentioned in the case, so you have to appoint three causes in your role-play, I think that is a good thing (40, f1)", or "I think that the criteria are realistic because the conversation has to be structured of course. You have to receive guidelines like 'you have to say something about this and about that' (92-93, f1)". Seniors, on the other hand, argued that "the big difference between the assessment and practice is that in the assessment you have to show all the criteria in 10 minutes, so you are focusing on ticking of all the criteria, while in practice, these criteria will come naturally during a conversation of half an hour (113, s3)", and "in practice you will also get to the point where everything is ok again and the client is satisfied, but you just get there in another way then you have to do here [at school] (9, s4)".

Moreover, seniors explicitly referred to the fact that having more experience in professional practice, changed how students dealt with assessments: "the more time you've been doing internships, in practice, the more experience you get and the better you possess all the skills and that is good for the assessments (177, s5)", and "the assessments deal with skills that we already possess, but I think that younger students cannot automatically carry out the assessments to a successful conclusion, that is a difference with us (128, s3)".

Conclusion and Discussion

This study compared two VET student groups, who differed in their amount of practical experience. It studied if the perceptions of assessment authenticity and alignment were stable and if the influence of the same kind of authentic assessment, or the perception thereof, on study approach and learning outcome (see Figure 7.1) was stable as well. Results showed that perception of authenticity and the influence thereof on student learning were fairly stable, with some salient differences. This finding has important implications for developing and using

authentic assessments in a curriculum, in which learning and working are integrated on a regular basis.

Overall, the *SEM* results supported the tenability and stability of the hypothesised model (see Figure 7.1). In line with our expectations, this meant that an increased perception of assessment authenticity resulted in more deep learning and development of generic skills in both groups. For freshman students, an increased perception of authenticity of all the assessment characteristics (task, physical context, form, and criteria) resulted in more deep studying and development of generic skills. For seniors this was also true for an authentic task, physical context and form. In terms of practical implications, this supports the use of authentic assessments during a VET curriculum and also suggests that not completely different kinds of assessments need to be developed for freshman and senior students.

However, some salient differences were found between freshman and senior students, which illustrate crucial elements that need to be taken into account in the operationalisation of authenticity for freshman or senior assessments. First, the *t*-test in this study revealed that freshman students perceived the assessment task and physical context as more authentic than senior students, and reported to have developed more generic skills in response to the assessment. Second, the *SEM* results showed that the assessment criteria had the strongest influence on deep studying and learning outcome for both groups, but that this influence was in opposite directions. Unexpectedly, an increase in perceived criterion authenticity resulted in *less* deep studying and development of generic skills for senior students.

Focus-groups data offered more detailed information on these findings and were examined to determine whether these differences could be explained by the fact that senior students have much more practical experience than freshman students. First, senior students referred explicitly to the fact that having more experience in practice made them more skilled to naturally deal with assessments of job performance. Specific assessment-criteria, even though perceived as being authentic, realistic steps taken in practice, inhibited seniors from performing the assessment naturally. It appears that seniors no longer need these analytic steps to successfully perform the assessment task. This was supported by Govaerts, Van der Vleuten, Schuwirth, and Muijtjens (2005) who found that senior students became demotivated by analytic criteria, while freshman needed analytic criteria to guide their learning. The "expertise reversal effect" (Kalyuga, Ayres, Chandler, & Sweller, 2003) argued that once students have gained expertise, they need less instructional guidance, because they have internalised the information. As a result of more experience with performing the assessment tasks in practice, seniors might perceive performing in practice differently (i.e., in a more integrated instead of a step-by-step way) than freshman students. Instructional guidance, in this case analytic performance criteria, becomes redundant for more experienced students and no longer contributes to their learning or even hinders it (Kalyuga et al.).

Senior students perceived the assessment task and physical context as less authentic than the freshman students. The qualitative results illustrate that seniors experienced the tasks as focusing too highly on general social-work activities. Seniors preferred more specialised instances, because they would learn more from them. It appears that the general tasks were only

'more of the same' for the more experienced senior students. Furthermore, seniors perceived the physical context, situated in an in-school role-play with a teacher, as redundant because of their experience on the work floor. They already performed these tasks in professional practice, which made them perceive performing these tasks in a role-play in school as less authentic. These results concerning the authenticity of the assessment task and physical context suggest that senior students already had experience with performing the tasks used in the assessments in professional practice. This experience possibly explains their feeling that they can perform the tasks in the assessments naturally, based on their experience. A combination of senior student perceptions of the (lack of) authenticity of the assessment task, physical context and criteria might explain why they reported developing fewer professional skills than freshmen in response to the assessment; seniors simply felt that there was less to learn because of their previous experiences in professional practice.

In short, when students (i.e., freshmen as well as seniors) perceive an assessment as more authentic, they report to study more deeply and to develop more professional skills. What students perceive as authentic depends on how they perceive professional practice and performance in professional practice (Gulikers, Kester, Kirschner, & Bastiaens, 2006; Lizzio & Wilson, 2004a; Messick, 1994). In addition, this perception of what professional practice and performing in practice is, can change when students gain more practical experience. This study suggests that more practical experience does *partly* change what students perceive as authentic and how an authentic assessment influences their learning. The quantitative findings reported in chapter 5, on the other hand, suggested that gaining practical experience during schooling (opposed to real practical experience after finishing school) did *not* change freshman and senior students' perceptions of authenticity. The additional qualitative data found in the current study, however, convincingly point out some salient differences with respect to what kind of operationalisation of the assessment dimensions freshman and senior students perceive as authentic and crucial for their learning. These data also show that when students do not perceive the assessment as appropriately reflecting professional practice, even if the assessment was developed to be authentic, it might not support, or even hamper, their learning. Taken together, the results of this study might mean that students with varying degrees of professional practice experience benefit more from different kinds of authentic assessment.

Practical Implications: Guidelines

For educational practice, at least for vocational types of education in which learning and working are intertwined, this means that using authentic assessment is useful and effective during a competency-based curriculum, but some critical issues need to be considered in the operationalisation for students with differing practical experience. Based on this study, students with more practical experience would learn more from assessments (1) that have holistic criteria that reflect *how* these students perceive performing in professional practice; (2) that use assessment tasks that reflect more specialised (out of the ordinary) professional activities instead of general professional activities practice or assessment tasks that allow students to tailor the tasks to their personal interests or working context, and (3) that are situated in real professional practice instead of in a role-play in school. Students with little practical experience, on the other

hand, prefer (1) more analytic criteria because performing in practice is also still a stepwise process for them and specific steps help them learn; (2) they are satisfied with assessment tasks that reflect more general professional activities; and, for them, (3) assessing in the workplace is not absolutely necessary, since assessing in a simulated or role-play setting can appropriate reflect performing in practice.

Moreover, based on the findings of this study, it is suggested that it is valuable to discuss with students what they see as authentic, how they perceive performing in professional practice, and how they think this professional practice should be represented in an assessment. In other words, more student involvement in the development of authentic assessments is recommended. This appears to be most critical with respect to the assessment criteria, since they seem to be the most influential element of an (authentic) assessment. When not operationalised correctly they can be detrimental for learning as was seen in the case of the senior students. When students are involved in the development of assessments for job performance, the chance that an assessment will be perceived as authentic by the students increases. As this study shows, when students perceive the assessment as authentic, they are stimulated deeper studying and the development of professional skills.

Limitations

First of all, this study and its implications are of relevance for schools where the goal of the assessments is to stimulate professional skill development or measuring of successful job performance and where learning and working are strongly integrated. With respect to the generalisability of the results of this study, three aspects need to be taken into account. First, VET is not current and used in the world, at least not in the way it is used in the Netherlands, especially with respect to the internship regimes. This might make it difficult to generalise outside of the VET context in the Netherlands. Moreover, this study examines the difference between students with little and much experience in professional practice and not between freshman and senior students in general. The results of this study cannot be transferred to the influence of authentic assessment on freshman and senior college students or higher education students, where learning and working are much less integrated. Second, this study was done by using one specific type of authentic assessment, namely a role-play. The results might be transferable to, for example, patient simulations in nursing or medicine, but not to completely other types of (authentic) assessments. Third, this kind of study should be replicated with other students groups, in other domains, and with other kinds of authentic assessments. As mentioned before, the study reported on in chapter 5 suggested that the perception of authenticity was completely stable throughout the educational career of VET students, while this study suggests that they are partly variable. The qualitative data favoured the idea of variability, but future research should be done to gain more insight.

In interpreting the results of this study it should be taken into account that because senior students have more experience in professional practice, they also have more experience with assessments at the work floor, while freshman students only have experienced authentic assessments in school. Not only previous experiences with professional practice, but also previous assessment experience might influence how students perceive an authentic assessment

and what kind of authentic assessment they need to stimulate their learning (Gulikers et al., 2004; Gulikers, Kester, Kirschner, & Bastiaens, 2006).

Future Research

This study showed that, in a vocational context, the relationships between student perceptions of authenticity, their study approach and professional skill development are rather stable, with some salient differences. This seemed to depend on student perception of what performing in professional practice involves. This has implications for using authentic assessment throughout a curriculum in which students gain more experience in professional practice, for example through internships. In addition, this study showed the added value of combining quantitative and qualitative data collection to get a clearer and richer picture of relationships and of reasons behind these relationships.

Future research should study authenticity perceptions and their influence on study approach and professional skill development in other educational contexts. Interesting contexts are types of education where learning and working are not as integrated or where the future work field is much broader and therefore less clear. In line with this, it would be interesting to study what kind of authentic assessment is most effective for student learning in the beginning of their educational trajectory in which they have little or no experience with working in professional practice and/or with authentic assessment. Professional development and assessment literature (Boshuizen et al., 2004; Kasworm & Marienau, 1997; Segers, Dochy, & Cascallar, 2003) advocate the use of authentic learning tasks or assessments early in the educational trajectory. However, previous studies suggested that these inexperienced students might have unrealistic perceptions of professional practice (Lizzio & Wilson, 2004b; Pena, 1997). The role of authentic assessment in this phase of an educational career might well be to help students create a more realistic idea of professional practice. The question then would be what kind of authentic assessment can give inexperienced students a realistic preview of professional practice. Future research should examine how the authenticity should be operationalised to stimulate the learning of these inexperienced and beginning students.

By comparing student perceptions about the authenticity of the same kind of authentic assessment, both within one group and between groups, this study gave indications about the important elements of assessment authenticity and how these should be operationalised to be effective for different student groups. The next step should be to compare assessments that do or do not take these elements into account and examine their impact on student learning and professional skill development. How do students perceive the authenticity of these assessments? What kind of study activities do they employ in response to the different assessments? The problem with these kinds of studies, and especially with conducting them in an ecologically valid setting, is that this often requires using and comparing different student groups, most likely from different schools, because one school does probably not have various versions of assessments. Since the impact of assessments is dependent on the whole learning environment (e.g., Biggs, 1993; Struyven, 2005) it might be difficult to distil the impact of the assessment. Thoroughly describing the research/school context and using qualitative data to examine the influence of a complex mix of factors is imperative in these cases (Birenbaum, 2003).

This study showed that perceived authenticity is an important element of new modes of assessment aiming at developing competencies or assessing job performance. Even though, authenticity is not the only criterion for valid assessment (Baartman, Bastiaens, & Kirschner, & Van der Vleuten, in press.; Dierick & Dochy, 2001; Linn, Baker, & Dunbar, 1991), we argue that an assessment that is perceived as authentic by students is an important step in the direction of bridging the gap between learning and working. Additionally, this study helps build this bridge throughout a curriculum in which learning and working are intertwined, by giving insight and guidelines for using authentic assessments that are perceived as authentic by students with different amounts of experience in professional practice.

Chapter 8

General Discussion

This final chapter combines the results of the studies described in the previous chapters in a reflection on the five-dimensional framework through the eyes of the different beholders who participated in these studies (i.e., freshman/senior students, teachers, and practitioners). In this reflection, the hypothesised relationships between previous experiences, beliefs, perceptions and learning are discussed. This reflection gives an insight into beliefs and perceptions of assessment authenticity of the different authentic assessment stakeholders and into the differences and similarities between the groups concerning the operationalisation of the five authenticity dimensions effective for learning. Practical guidelines and rules of thumbs for developing authentic assessments are distilled. In the end, several critical remarks and suggestions for future research are outlined.

"Authenticity is in the eye of the beholder" means that whether an assessment is perceived to be authentic or not depends on who is looking at it. Authentic assessment aims at bridging the gap between learning and working by focusing on assessing successful job performance and stimulating students towards deep studying and professional skill development. However, as authenticity is in the eye of the beholder, it is imperative that those being assessed – the students - perceive the assessment as being authentic with respect to professional practice, before it will stimulate their learning. It was hypothesised that when students perceive an assessment as more authentic, they are stimulated to deeper studying and professional skill development (see Figure 8.1). Before this could be studied, we needed to gain insight into what assessment characteristics were important for assessment authenticity. Therefore, a literature review was conducted that resulted in a five-dimensional framework describing five assessment characteristics determining authenticity. The authenticity of an assessment was defined by the degree of resemblance between these five assessment characteristics and the professional practice situation the assessment aims to reflect, at the educational level of the student (i.e., the criterion situation). Additionally, it was hypothesised that previous experiences in professional practice or previous experiences with authentic assessments influence a person's beliefs about what authentic assessment is. These beliefs, in turn, were hypothesised to influence that person's perception of the authenticity of a newly encountered assessment (see Figure 8.1).

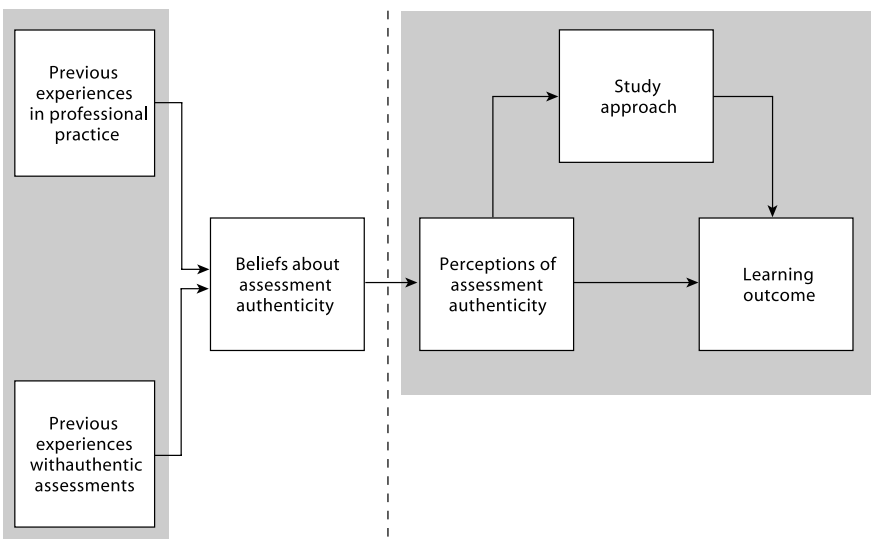


Figure 8.1. Relevant variables for authentic assessment and their relationships

Beliefs and perceptions concerning the identified assessment characteristics of the 5DF were compared between groups that differed in their amount of experience with working in professional practice and/or with authentic assessments. These groups represented different

stakeholders (i.e, student, teacher, and practitioner) as well as different student groups (i.e, freshman and seniors). Table 8.1 gives an overview of the participant groups in the different studies and the variables that were examined.

Table 8.1. An overview of the studies, participants and variables

Chapter	Methodology	Field	Variable	Participants
2	Qualitative/ Quantitative	Nursing	Beliefs	Freshman (LEPP, NEAA) and Senior students (MEPP, EWA) following a Vocational Training Programme
			Perceptions	Freshman (MEPP, NEAA) and Senior students (MEPP, EWA) following a Block-day release Training Programme
				Teachers
3	Quantitative	Social Work	Perceptions	Freshman students (LEPP, NEWA²)
				Teachers
4	Qualitative/ Quantitative	Nursing	Beliefs	Freshman students (LEPP, EWA)
				Teachers
				Practitioners
5	Quantitative	Social Work	Perceptions	Freshman students (LEPP, NEWA²)
				Senior students (MEPP, EWA)
				Two teacher groups
6	Quantitative	Social Work	Perceptions	Senior students (MEPP, EWA)
			Study approach	
			Learning outcome	
7	Qualitative/ Quantitative	Social Work	Perceptions	Freshman students (LEPP, NEWA)
			Study approach	Senior students (MEPP, EWA)
			Learning outcome	

Note. LEPP = Little experience in professional practice; MEPP = More experience in professional practice; NEAA = No experience with authentic assessments; NEWA = No experience with workplace assessment, but experience with authentic assessments in school; EWA = Experience with workplace assessments

Many similarities were found between the beliefs and perceptions of different groups concerning the important elements of authentic assessment and their influence on learning. However, some salient differences were found as well. These differences and similarities pointed out what the determining facets of authenticity are as well as how these facets should be operationalised for different student groups to be perceived as authentic and as a result to stimulate student learning.

This final chapter reviews the results in a topical instead of chronological manner by reflecting on the five-dimensional framework (5DF) through the eyes of the different beholders (see Table 8.1). We chose to organise this discussion this way, because we feel that comparing

beliefs and perceptions of different groups per dimension gives the reader the best insight in the differences and similarities with respect to the five dimensions of assessment authenticity and the implications for developing authentic assessments that are effective for stimulating learning for different student groups. In addition, this gives the best insight into the relationships between the variables hypothesised in Figure 8.1. Practical implications with respect to the five dimensions are directly combined in the review of the results and in the end, some overall practical implications or guidelines are described. In the last section of this chapter several critical remarks of the study and suggestions for future research are outlined.

Reflections on the Five-Dimensional Framework

Task

The task seems to be one of the most important assessment characteristics for providing an authentic experience [2, 4, 6, 7][7]. Moreover, there was a relatively strong and positive influence of perception of task authenticity on deep studying and generic skill development, which seemed to be stable between groups [6, 7]. In addition, perception of task authenticity showed a positive relationship with student grades as well [6].

Most data supported that having more experience in professional practice or having experience with authentic assessment (both in school or at the workplace) influenced both student beliefs of what constitutes an authentic task [2, 4], as well as their perception of the authenticity of an assessment task [2, 7]. Only the data concerning both student groups in chapter 5 and the Vocational Training Programme (VTP) and Block Release Programme (BRP) student-groups in chapter 2 did not support this influence. Qualitative supporting data showed that experienced students agreed that an authentic task that aims at stimulating professional skill development, should not be completely pre-structured, because dealing with the element of surprise is representative of real work-situations. Students who either have had a lot of experience in professional practice, or students who have had experience with work-based authentic assessment, argued that an authentic task should relinquish some ownership to the students so that it allows them to adapt the task to their own interest or problems they personally encounter during their work placements [4, 7]. This also assured a complexity level that is representative of professional practice at the level at which the student has to perform (i.e., during internships) [2, 4]. On the other hand, students who have had little experience in professional practice and no experience with authentic assessment [2] focused more on assessment tasks that are relatively structured, easy, and mainly directed to knowledge instead of an integration of knowledge, skills and attitudes. It can be argued that these students still had relatively traditional beliefs about assessment that guided their perception of the authenticity of newly encountered assessments. The findings corroborated, for the most part, the hypothesised influence of previous assessment and professional practice experiences on student beliefs and perceptions of task authenticity.

[7] The numbers between brackets refer to chapter numbers.

Comparing the beliefs and perceptions of teachers, students and/or practitioners shed an interesting light on what constitutes an authentic task. They agreed that such a task should deal with representative work activities and competencies and should focus on performance. Interestingly, however, tasks that were authentic in the eyes of the teacher were not automatically authentic in the eyes of practitioners and/or students [2, 4, 5]. Two reasons for this finding could be given based on the findings in this thesis. First, teachers are not always up-to-date with developments in professional practice [4]. Second, teachers are the ones to develop the assessment and as a result perceive the authenticity of their own developed assessment as highly authentic [5].

For educational practice, these findings suggest that teachers should not try to make the assessment completely authentic in the eye of the students. They should describe the task in terms of professional competencies and work activities, but allow students to tailor them to their own interests or internship context. However, this might not hold for students who have both very little professional practice experience and no experience with authentic assessments. These students are not focused (yet) on developing assessments that resemble their future professional practice. Instead they are guided by traditional assessment beliefs shown by their emphasis on structured, easy and knowledge-based tests. These students probably need extra guidance when confronted with authentic kinds of assessment. We suggest that this guidance focuses on helping these students to translate a general assessment task to their own internship situations, interest, or learning goals instead of on providing these students with assessment tasks that are made authentic for them.

Physical context

The physical context is a very obvious facet of authenticity in the eyes of all participants, indicated by the high reliability of the physical context scale in the original [3] and revised versions of the authenticity perception questionnaire [5, 6, 7]. Moreover, we found a positive influence of perceived physical context authenticity on deep studying and generic skill development [6, 7]. However, the authenticity of the physical context seemed to be of differing importance to different student groups, and, contrary to the theories of situated learning (Brown, Collins, & Duguid, 1989; Lave & Wenger, 1991), the physical context did not seem to be the most important facet of authenticity. Both student and teacher perceptions of the role-play assessment showed that an assessment, and the other dimensions of the 5DF, could be valued as fairly authentic, even though the physical context was not perceived as very authentic [5]. This is in line with previous research (Gulikers, Bastiaens, & Martens, 2005; Van Merriënboer, 1997) that argued that an authentic physical context is less important when both the assessment task and form are authentic. These results corroborate previous studies that argued that an authentic physical context (i.e., in the workplace) is an important but not sufficient condition for authentic assessment (Messick, 1994; Stein, Isaacs, & Andrews, 2004).

However, the importance of a more authentic physical context seemed to increase when students gained more insight in professional practice [2, 4, 7]. Students with little experience in professional practice or without experience with assessments at the workplace saw that assessing at the workplace is different from assessing in school, but they believed that school assessments

could be appropriate for assessing job performance [2, 7]. Students with more experience in professional practice or with experience with assessments at the workplace felt that simulations or role-plays were valuable and even crucial for preparing students for the workplace, but that summative assessment of job performance needed to be done in the workplace. They perceived summative assessment in school as fake or redundant. Also teachers and practitioners said that simulations and role-plays in school were crucial for learning, but that summative authentic assessment should be done at the work floor [2, 4]. These differences between participant groups again supported the hypothesis that previous experiences influence beliefs about authentic assessment.

In terms of practical implications, these findings suggest two things. First, for students with little experience in professional practice and no experience with assessment in the workplace, an authentic assessment does not have to take place in the workplace. In their eyes, assessment of job performance can well be done in school, as long at the assessment task focuses on professional activities. The importance of using authentic workplace assessments should increase when students get a better idea of what professional practice looks like. Second, the finding that practicing with performing authentic tasks in school is univocally perceived as crucial for learning, stresses the importance of integrating instruction, learning and assessment where learning and assessment tasks both reflect authentic whole tasks (Birenbaum et al., 2006; Biggs, 1996; Van Merriënboer, 1997). The main point is that, during the curriculum, students should be allowed to practice performing the whole task representative of the profession in several situations before they are going to be formally assessed on performing this task.

Social Context

The social context seems to be the least important characteristic of authenticity. Students and teachers rated it as the least important dimension of the 5DF [2], they did not discuss it much [2, 4] and the social context scale did not reach an acceptable reliability level for students both in the original [3] and the revised perception questionnaire [5, 6, 7]. The qualitative data suggested that teachers in the field of nursing believed assessment to be an individual affair, even though they realised that practice often required collaboration [2, 4]. Practitioners in this field, on the other hand, argued that letting students collaborate with colleagues at work during an assessment is both authentic and valuable for learning [4]. In their view, only using individual assessments was not authentic, since working in teams is inextricable bound up with the nursing job of today. This suggested that teachers were not really guided by current professional practice, but rather by beliefs based on traditional school situations where assessment is purely individual. Students perceived professional practice as requiring a lot of teamwork, and collaboration was seen as a very important competency, but most students had never considered the idea of collaborative assessment [2, 4].

At this moment, it looks like teachers and students are not yet ready for collaborative assessment. We would not want to conclude that the social context is not a dimension of assessment authenticity, because all stakeholders recognise that individual as well as collaborative activities are representative of professional practice, but considering the social context in developing authentic assessment is not yet seen as important.

For educational practice, this suggests that when a school is just at the beginning of developing and implementing authentic assessments, time, money and energy can be saved on the social context as this is seen as the least important element of authentic assessment, at least up till now. When schools want to change towards more collaborative assessment, starting with changing traditional assessment beliefs might be an important first step. Going against current beliefs will always cause resistance (Gibbs, 1992) and Van Rossum and Hamer (2003) argued that when we want to innovate education, we need to explicitly address current beliefs and help them to change to fit the new educational ideas.

Form

The form of the assessment seems to be important for authenticity [2] and it needs to be represented as a separate dimension of assessment authenticity [3]. In addition, perceived assessment-form authenticity had a strong positive relationship with student grades [6] and was the only assessment characteristic that had a *direct* significant impact on generic skill development, without influencing study approach [6, 7].

The perception of the authenticity of the assessment form seemed fairly stable [5, 7]. When different student groups were confronted with the same form of authentic assessment, they agreed in their perception of its authenticity. However, when directly asking students what they believed to be an appropriate form of an authentic assessment for assessing job functioning, some differences appeared. For example, freshman nursing students without either professional practice or authentic assessment experience [2] believed that a realistic case description was an authentic assessment of job functioning. Other, more experienced students stressed the need for performance-based assessments as role-playing or observations on the work floor. The major difference between these student groups was that the former group had only had experience with 'traditional' education as learning and testing in school. This might have caused these students to have a more narrow view on authentic assessment as compared to students who already encountered more authentic assessments, or students who have had more experience with performing in professional practice. This finding again supported the hypothesised influence of assessment and professional practice experience on beliefs and perceptions of authenticity.

Practitioners and students with professional practice or authentic assessment experience agreed that an authentic assessment of job performance required multiple assessment moments, multiple assessment methods and formative assessments that allowed students to practice the things they needed to do in the summative assessment. Nursing students and practitioners of the study described in chapter 4 gave a more detailed insight in this issue. They argued that an authentic assessment of job performance needs to entail a combination of assessments, because job performance entails several aspects that cannot be assessed authentically with one single method. They argued that successful functioning in a job, at least nursing, involves generic skills (e.g., communication, planning, working in a team), technical skills (e.g., giving an injection or applying a bandage) and knowledge. To assess these three aspects authentically (i.e., assess if they are used as required in practice) favours the use of different assessment methods for the different aspects of job performance.

The knowledge aspect deserves some extra attention, because teacher groups argued *against* separate knowledge testing [2, 4], while several student groups and practitioners argued *for* incorporating knowledge testing in an authentic assessment [2, 4]. The reasons for wanting knowledge testing as part of an authentic assessment, however, differed between student groups. Here, we again saw a discrepancy between students without authentic assessment experience and little practical experience, and the other student groups. The former group stressed knowledge testing in the 'traditional' sense because this was the kind of testing they were familiar with. On the other hand, students with more experience in professional practice, as well as practitioners, stressed the need for knowledge testing because they perceived knowledge as a fundamental element of competent performance [4]. They argued that in order to assess whether a student is capable of successful job performance or to stimulate students to learn the things relevant for practice, an authentic assessment should involve explicit testing of "why" knowledge. This testing should show that students know the reasons for and the consequences of performing in a certain way. The ideas of the latter groups are more in line with current ideas about assessment of competencies in which not only the use of multiple assessment moments and methods is stressed, but also the idea of combining new and old (i.e., traditional) methods of assessment to appropriately assess competencies (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006; Segers, Dochy, & DeCorte, 1991; Straetmans, Sluijsmans, Bolhuis, & Van Merriënboer, 2003; Van der Vleuten & Schuwirth, 2005).

For educational practice, these results suggest that an authentic assessment requires the use of multiple assessment moments and assessment methods. We, supported by students, practitioners and current views on assessment, suggest the use of different methods for different aspects of job performance (e.g., technical skills, generic skills, or knowledge). This is important for both gaining a complete as possible a picture of a student's capability to function in an job, as well as for stimulating students to learn all the aspects needed for competent job-performance. Generic skills require observations in professional practice during a longer period of time and technical skills can be assessed at fixed moments, but preferably more than ones. Knowledge testing as part of an authentic assessment should be directed towards applying knowledge and knowing "why". This means that assessment of knowledge should be directly connected to performing (e.g., as part of a performance assessment), and should not be treated as an isolated activity.

Criteria

Both quantitative and qualitative analyses showed that the criteria of the assessment had the strongest influence on student learning and skill development [4, 6, 7]. Practitioners argued that students only do what is in the criteria [4] and students said that what they study is mainly guided by what is in the criteria [7].

However, based on the results of this thesis it can be said that developing authentic assessment criteria is easier said than done, as several issues complicate the development of authentic criteria that positively influence student learning. In the eyes of students and practitioners, school criteria often did not appropriately reflect professional practice for three reasons [2, 4, 5, 7]. First, teachers (i.e., the developers of the assessment) were often not up-to-

date with respect to developments in professional practice, which was reflected in the assessment criteria they developed. Second, schools placed too much emphasis on technical skills at the expense of generic skills, while these have become the primary requirements for professionals (Lizzio & Wilson, 2004b; Onstenk, 1997; Semeijn, 2005). Third, students and practitioners often perceived the criteria to reflect school requirements and procedures instead of requirements and procedures used in professional practice (one practitioner [4], for example, said "protocols that we use here are not synonymous to what students learn in school … but the student should be able to use the protocol that is used in this work setting"). Following Messick's line of reasoning (1996), it could be argued that the criteria might have been authentic with respect to the school objectives, but not with respect to professional practice, because the school objectives did not reflect professional practice.

In sum, these three issues could lead to assessment criteria that were not representative of professional practice, at least in the eyes of students and practitioners. As a result, what students learnt for such an assessment, was not representative of professional practice either [4, 7]. Thus, authentic criteria seem to be crucial for authentic learning.

Two practical guidelines can be described based on these findings. First, to assure that the assessments reflect up-to-date professional requirements, teachers need to be stimulated to keep up-to-date with developments in professional practice and practice should be more involved in the development or interpretation of assessment criteria (Kerka, 1995). Second, to increase the likelihood that students perceive the assessment criteria as being authentic, and thereby stimulating for their learning, authentic criteria should give students the possibility to translate the criteria to their own working context (e.g., workplace, institute or traineeship). The criteria should describe guidelines and set preconditions, but at a level that is general enough to allow students, with or without the help of practitioners, to tailor the assessment to their internship context (Klarus, 2000; Tigelaar, 2005).

How authentic criteria should be operationalised to stimulate student learning depends first on what kind of output the assessment aims to assess [4] and second on how students perceive performing in professional practice [7]. With respect to the first factor, students and practitioners believed that different kinds of criteria should be used for different kinds of outputs that are assessed. This corroborated the refinement of the criterion dimension based on the results reported in chapter 3 that suggested that authentic criteria should be directly related to job-relevant results. More concrete, students and practitioners believed that analytic stepwise criteria are authentic for assessing technical nursing skills, but not for generic professional skills such as communication, collaboration, or taking initiative [4]. Authentic criteria for generic skills are holistic, interpretative criteria against which student performance can be judged by several involved parties like colleagues of the student at place of internship, mentors and the student him/herself. This is in line with the ideas of Guba and Lincoln (1987) who argued that interpretation from and communication between different involved parties gives the most adequate (i.e., authentic) picture of a students competencies.

With respect to the second factor, the quantitative data showed that the relationship between perception of criterion authenticity and deep studying was not stable between groups that

differed in the amount of practical experience that they have had [7]. The qualitative data showed that students with differing degrees of experience in professional practice stressed different operationalisations of authentic assessment criteria. Analytic criteria positively influenced deep learning of inexperienced freshman students, while these same criteria had a negative influence on deep learning of more experienced seniors. In the eyes of freshman students, the analytic criteria were authentic with respect to *what* steps describe a performance in practice and *how* these steps should be performed in practice, namely step-by-step. Seniors, on the other hand, perceived the analytic criteria as describing realistic steps to be taken in performance in practice, but in their eyes, these criteria did not realistically describe *how* the task is performed in a natural professional situation (one student, for example, stated that "you use all the steps in the criteria in practice, but they will come naturally in a 30 minute conversation with a client"). These results suggested that especially the perception of the "how-part" changed when students gained more experience in professional practice, because the perception of how tasks are performed in professional practice changed. It seems that more experienced students did not perceive performing in professional practice as a step-by-step process anymore. In line with some results from studies on developing expertise (Kalyuga, Ayres, Chandler, & Sweller, 2003; Van Merriënboer, 1997), we could argue that these students had internalised the performance steps and chunked them into more integrative wholes. As a result, learning of more experienced students can be hampered when instruction (in this case the assessment criteria) describes performing in an step-by-step manner, while these step-by-step criteria are beneficial for initial learning. These results again supported the hypothesis that prior practical experiences influence what students believe and perceive as being authentic, which in turn, influence their learning.

In terms of practical implications, these results suggest that authentic criteria not only need to reflect *what* is used in practice (knowledge, skills and attitudes) but also *how* it is used to perform in the professional practice situation. Moreover, both "what" and "how" criteria should describe performing in professional practice as students perceive this. Results suggested that the perception of the "how-part" changes when students gain more experience in professional practice, implying that especially these criteria need to change during a curriculum in which learning and working are alternated on a regular basis. Students with little experience in professional practice learn more from specific, analytic, step-by-step criteria. More experienced students, on the other hand, benefit more from holistic criteria that mainly show what has to be the end result (e.g., the client needs to be satisfied with the discussed solution) and allow students to think of their own way to come to this result. By incorporating both "what" and "how" criteria that reflect performing in professional practice in the eyes of the student, assessment might be more able to guide both *what* students learn, as well as *how* they learn (Scouller & Prosser, 1994; Van Gog, Paas, & Van Merriënboer, submitted).

Criterion Situation

If we take all these reflections back to our theoretical framework described in chapter 2, then the idea of the *criterion situation* is corroborated. To recapitulate, the criterion situation represents the professional practice situation on which the authentic assessment is based, with the extension

that this situation should be described at the educational level of the student. This means that an authentic assessment should not by definition reflect the professional practice situation at expert level, but that it should reflect the professional practice situation as students at their educational level see and experience it (e.g., in work placements). As these results show, this is not necessarily the same as what professional practice actually is. Moreover, it seems to be different for student groups with differing kinds of professional practice or assessment experiences.

General Practical Implications: Guidelines

Several *general* practical guidelines for using authentic assessments during a curriculum can be described based on the findings concerning the influence of previous assessment and professional practice experiences on beliefs and perceptions of assessment authenticity:

Confront Students with Authentic Assessments Early in their Educational Trajectory.
These assessments do not have to reflect maximum authenticity from the start. In the beginning, or when dealing with students without authentic assessment experience and with little or no professional practice experience, assessments such as realistic case-based assessments might be appropriate. These kinds of authentic assessment, with authentic tasks but relatively inauthentic other dimensions of authenticity, can be helpful to confront traditional beliefs of inexperienced students. Then, in-school assessments with authentic tasks and an authentic form, such as role-playing with realistic case descriptions as starting point, seem to be helpful for giving students with little experience in professional practice a preview of what working life looks like (Gulikers et al., 2005). This is an important start towards bridging the gap between learning and working (Boshuizen et al., 2004). At this point, the assessment criteria should be fairly analytic because students do not yet have a holistic view of performing in professional practice. When students gain more experience with performing in professional practice and with workplace assessments, authentic assessments situated at the workplace and with more holistic criteria become relevant for their learning.

Explicitly Communicate the Authenticity of a Certain Assessment and Create Mutual Understanding between Involved Stakeholders.
This step is often forgotten, as "everybody thinks that we know what we are talking about" (Petraglia, 1998). However, this thesis showed that different stakeholders have different opinions about what an authentic assessment should look like, which makes explicating of major concern for developing authentic assessments that are beneficial for student learning as well as valid for assessing job performance (Kerka, 1995). When different stakeholders (i.e., teachers, students, and practitioners) are involved in the development of the assessment, these parties should sit down to make their beliefs explicit before starting the development process. When students and/or practitioners are not involved in the development, but only in the use of the assessment, the developers should both communicate the link between the assessment and working life (Lizzio & Wilson, 2004a) as well as help students to understand how the, often theoretically described, assessments (e.g., competencies or criteria) can be translated to concrete actions or activities that are relevant for practice.

Educational Practice can Benefit from Using the 5DF.

It is a tool for talking and thinking about authentic assessments, for making implicit beliefs of various involved stakeholders explicit, for developing assessments with different kinds or amounts of authenticity, for evaluating the authenticity of the assessment in the eyes of users, and to study the influence of these assessments on learning and competency development. There is not one best authentic assessment. We should be aware that different student groups stress different operationalisations of the five dimensions for the assessment to be best for their learning. The 5DF offers opportunities to develop different kinds of authentic assessments during a curriculum.

Table 8.2 summarises both the general practical guidelines as well as the rules of thumb described in the reflection of the five dimensions of the 5DF.

Table 8.2. Practical guidelines and rules for developing authentic assessments

General guidelines
Confront students with authentic assessment early in their educational trajectory.
Explicitly communicate the authenticity of a certain assessment and create mutual understanding between involved stakeholders.
Use the 5DF for thinking about and explicating ideas concerning authentic assessments, and for developing and evaluating authentic assessments.
Rules of thumb concerning several authenticity dimensions
Integrate instruction and assessment by offering opportunities to perform authentic, integrated tasks(i.e., learning tasks/formative assessments) in and out of school to prepare students for summative authentic assessment.
Stimulate teachers to keep up-to-date with developments and requirements in professional practice.
Allow student to tailor the assessment task and criteria to their own situation (work context, interest, learning goals).
Rules of thumb concerning specific authenticity dimensions
Do not make the assessment *task* completely authentic *for* students, but help them to make the task authentic for themselves.
Increase the authenticity of the *physical context* as student gain more experience with working or assessing in practice.
Leave the *social context* alone when just starting to use authentic assessments.
When considering an authentic *social context*, first deal with traditional beliefs.
An authentic assessment *form* should involve multiple assessment methods and moments for different aspects of job performance.
Think about incorporating knowledge-testing directed at *knowing why* as part of the authentic assessment *form.*
Involve practice in the development and interpretation of authentic assessment *criteria*.
Authentic assessment *criteria* should deal with *what* is used in practice, as well as with *how* this is used. How-criteria should develop from being step-by-step to being more holistic as students gain more experience with performing in practice.

Concluding Remarks

Several indicators were found to support the hypothesis that when students perceive an assessment as being more authentic, they are stimulated to study more deeply and to develop generic skills. This is an important and promising finding, because even though all new modes of assessment aim at stimulating deep studying and thereby improving learning outcomes, empirical findings that support the relationship between assessment characteristics and a deep study approach and between a deep study approach and improved learning outcome are scarce (Marton & Säljö, 1997; Ramsden, 1992).

In addition, the research resulted in several findings supporting the hypothesis that the variables 'amount of experience in professional practice' and 'previous experience with authentic assessments' influence authenticity beliefs and perceptions and, thus, also influence what kind of authentic assessment different students need to stimulate their learning.

Critical Remarks and Directions for Future Research

Several critical remarks to this study need to be addressed that directly lead to a number of issues to be attended to in future research.

This thesis dealt with differences in perceptions of authenticity, how this was affected by experience, and how it affects studying. It dealt not with the authenticity of different kinds of assessment. To this end we chose role-playing as an example of an authentic assessment and examined variations in perceptions and the influence thereof on studying between and within groups. Whether the findings are generalisable to other authentic assessments remains to be seen. Future research should, thus, examine variations in student perceptions of different kinds of authentic assessment and the impact of these different authentic assessments on student learning. Important in this respect is that the five-dimensional framework offers possibilities to classify assessments with varying degrees or kinds of authenticity, which offers opportunities for more controlled comparison between different kinds of assessment on student learning.

Another limitation of this study is that the influence on *student learning* was, to a considerable extent, examined via *self-report questionnaires*. Even though the scales of the study approach and the generic skill development questionnaires were reliable, self-report questionnaire have been criticised for not being predictive of actual behaviour (Biggs, Kember, & Lueng, 2001; Lizzio, Wilson, & Simons, 2002). In addition, the qualitative data that were used to complement the questionnaire data, were self-report data as well. Students discussed in interview sessions what they did in response to an assessment or what they would do when confronted with other kinds of assessment. To gain more insight in the relationships between assessment and student learning, future research should also study student learning by monitoring, observing or logging the kind of learning activities students employ when confronted with different assessments.

Another issue with respect to using questionnaires is that results of chapter 3 revealed that the *reading level* of respondents, can colour the results of questionnaire data. When a questionnaire requires a reading level that is above the reading level of the students, the chance of reaching reliable scales decreases. Future research should take advantage from considering the

reading level of participating students. Readability measurements can be used when developing questionnaires to check the readability before testing or using the questionnaire in context.

The model presented in Figure 8.1 presupposes both interdependence and development. By interdependence we mean that different elements in the model affect other elements. By development we mean that the different elements change across time. This thesis assessed several elements of this model, but did not assess the model as a whole. Future research should involve longitudinal studies that make it possible to assess all the variables within one developing group of students throughout their education. This also allows for the examination of whether changes in one part of this whole model result in changes in the following parts as well. Based on the experiences we have had and the results of the studies, we would suggest a combination of quantitative and qualitative data-collection to gain the best insight into the variables and especially the relationships between the variables. Structural equation modeling, as used in this thesis, seems to be promising for testing relationships and the hypothesised causalities between variables. However, when research is done in an ecologically valid setting, the relationships between all involved variables will be too complex to catch with only quantitative data (Birenbaum, 2003).

Another issue has to do with the contextualisation of this study in Vocational Education and Training (VET) in the Netherlands. This *context* was chosen because of its fitness to our research purposes and to our definition of authenticity, since the primary goal of this type of education is to prepare students for the labour market. However, it is possible that the results of this study will not be valid in other types of education where there is only a vague notion of what the future labour market entails, where both education and students are more theoretically-oriented, or where students are not given the possibility to carry out internships while studying. However, authentic assessments are also an issue in these less work-oriented types of education. There are, for example, signs that students in higher education or adult- and distance-education, also perceive a lack of fit between what they are studying and what they will be doing in their future work (Hager & Hofgard-Lycke, 2005; Semeijn, 2005). These students prefer for their assessments to be relevant for their future lives as well (Huang, 2002; Newmann, 1997; Pena, 1997; Smith & Koshy, 2005). This also appears to be the case in primary education, where authentic assessments are thought to be important for engaging and motivating pupils (Henderson & Karr-Kidwell, 1998). Though for these other groups of students, defining authenticity by its resemblance to students' future professional practice might not be the best operationalisation. Future research, thus, should examine what authenticity means in other types of education and how assessments need to be operationalised to be perceived as being authentic in the eyes of their students.

Even though authenticity is seen as crucial element of new modes of assessment, it is not the only quality criterion for valid competency-assessment (Baartman et al., in press; Dierick & Dochy, 2001; Linn, Baker, & Dunbar, 1991). Authenticity needs to be considered in relation to other criteria, because increasing the authenticity of an assessment could mean making concessions to other quality criteria such as comparability or reproducibility. Questions that still remain are: What are the consequences for other quality criteria when the authenticity is

increased? And what kind of concessions to other quality criteria are allowed in valid assessment practices and when are concessions with respect to authenticity needed?

A final issue that needs to be discussed is the relationship between authenticity and *student motivation*. In this thesis, we referred several times to the expected positive effect of authenticity, or student perception thereof, on student motivation. However, none of the studies took this motivational aspect into account. This thesis showed that students who perceived the assessment as being more authentic, compared to other students, studied deeper and developed more generic skills. It can be argued that this finding is due to the fact that when students perceive an assessment as more authentic, their motivation to learn increases, which results in higher learning outcomes (Herrington & Herrington, 1998; McDowell, 1995). However, previous research showed that students who were more motivated when working with an authentic task, did not do more, rather they tended to do different things (Martens, Gulikers, Bastiaens, 2004), while in our research, students reported doing more. Future research should study the relationships between perception of assessment authenticity, motivation and student learning. Does an authentic assessment first have to stimulate student motivation, before it will positively influence student learning? Is this relationships stronger for students who lack intrinsic motivation than for students who are inherently more intrinsically motivated? Or do students who lack intrinsic motivation benefit more from authentic assessment that intrinsically motivated students? Future research should take the impact of authentic assessments on student motivation into account and, in turn, examine not only how much these students do, but also what kind of learning activities they employ.

References

Alderson, C., & Wall, D. (2003). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

Alessi, S. M. (1988). Fidelity in the design of instructional simulations. *Journal of Computer-Based Instruction, 15*(2), 40-47.

Arter, J. A., & Spandel, V. (1992). An NCME instructional module on: Using portfolio of student work in instruction and assessment. *Educational Measurement: Issues and Practice, 11*(1), 36-45.

Baarda, D. B., de Goede, M. P. M., & Teunissen, J. (2001). *Basisboek kwalitatief onderzoek: praktische handleiding voor het opzetten en uitvoeren van kwalitatief onderzoek* [Basic book qualitative reserach: practical manual for desinging and performing qualitative research. Houten: Stenferd Kroese.

Baartman, L. K. J., Bastiaens, Th. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (in press). The wheel of competency assessment: presenting quality criteria for competency assessment programmes. *Studies in Educational Evaluation, 32.*

Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum Associates.

Biemans, H., Nieuwenhuis, L., Poell, R., Mulder, M., & Wesselink, R. (2004). Competence-based VET in the Netherlands: Background and pitfalls. *Journal of Vocational Education and Training, 56*, 523-538.

Biggs, J. B. (1987). *The Study Process Questionnaire Manual*. Melbourne: Australian Council for Educational Research.

Biggs, J. B. (1989). Approaches to the enhancement of tertiary teaching. *Higher Education Research and Development, 8*, 7-25.

Biggs, J. B. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347-364.

Biggs, J. B., Kember, D., & Leung, D. Y. P. (2001). The revised two-factor Study Process Questionnaire: R-SPQ-2F. *British Journal of Educational Psychology, 71*, 133-149.

Birenbaum, M., & Dochy, F. (1996). *Alternatives in assessment of achievements, learning processes and prior knowledge*. Boston, MA: Kluwer Academic Publishers.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3-29). Boston: Kluwer Academic Publishers.

Birenbaum, M. (2003). New insights into learning and teaching and the implications for asessment. In M. Segers, F.. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards*. Dordrecht: Kluwer Academic Publishers.

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgeway, J., & Wiesemes, R. (in press). A learning integrated assessment system. Position paper of the Special Interest Group Assessment of the European Association of Research in Learning and Instruction. *Educational Research Review.*

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Bloom, B. S. (1956). *Taxonomy of educational objectives. The classification of educational goals*. London: Longman Group Limited.

Boshuizen, H. P. A., Bromme, R., & Gruber, H. (2004). *Professional Learning: Gaps and transitions on the way from novice to expert*. Dordrecht: Kluwer Academic Press.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15*, 101-111.

Boud, D. (1995). Assessment and learning: contradictory or complementary? In P. Knight (Ed.), *Assessment for learning in higher education* (pp. 35-48). Londen: Kogan Page.

Boud, D. E. (1998). *Current issues and new agendas in workplace learning*. Leabrook, South Australia: National Centre for Vocational Educational Research Ltd.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*, 32-42.

Brown, S., & Knight, P. (1994). *Assessing learners in higher education*. London: Kogan Page.

Byrne, B. M. (2001). *Structural Equation Modeling with AMOS. Basic concepts, applications and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Calderón, J. L., Hays, R. D., Lui, H., & Morales, L. S. (2005). Variation in readability within surveys. Retrieved February 10, 2005, from Center for Health Improvement for Minority Elders web site: http://www.chime.ucla.edu/measurement/presentations/

Cattell, R. B. (1966). The scree test for numbers of factors. *Multivariate Behavioral Research, 1*, 245-276.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hilldale, NJ: Erlbaum.

Cooper, B. (1994). Authentic testing in mathematics? The boundary between everyday and mathematical knowledge in national curriculum testing in English schools. *Assessment in Education: Principles, Policy & Practice, 1*(2), 143-166.

Cronin, J. F. (1993). Four misconceptions about authentic learning. *Educational Leadership, 50*(7), 78-80.

Cummings, J. J., & Maxwell, G. S. (1999). Contextualising authentic assessment. *Assessment in Education: Principles, Policy & Practice, 6*(2), 177-194.

Dall'Alba, G., & Sandberg, J. (1996). Educating for competence in professional practice. *Instructional Science, 24*, 411-437.

Darling-Hammond, L. (1994). Setting standards for students: The case for authentic assessment. *The Educational Forum, 59*, 14-21.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment in teaching in context. *Teaching and Teacher Education, 16*, 523-545.

Dierick, S., & Dochy, F. (2001). New lines in edumetrics: new forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*, 307-329.

References

Dochy, F. (2001). A new assessment era: different needs, new challenges. *Research Dialogue in Learning and Instruction, 10*, 11-20.

Dochy, F., & Moerkerke, G. (1997). Assessment as a major influence on learning and instruction. *International Journal of Educational Research, 27*(5), 415-431.

Dochy, F., & McDowell, L. (1998). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*, 279-298.

Drew, S. (2001). Student perceptions of what helps them learn and develop in higher education. *Teaching in Higher Education, 6*(3), 309-331.

Dunteman, G. H. (1989). Principal component analysis. Thousand Oaks, CA: Sage Publications.

Entwistle, N., McCune, V., & Hounsell, J. (2002). *Approaches to studying and perceptions of university teaching-learning environments: Concepts, measures and preliminary findings* (Occasional Report 1). Retrieved March 6, 2005, from http://www.ed.ac.uk/etl.

Entwistle, N. J., & Ramsden, P. (1983). *Understanding Student Learning*. London: Croom Helm.

Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment. Introduction to the special issue. *Higher Education, 22*, 201-204.

Eurydice (2004). The educational system in the Netherlands 2004 (Dutch Eurydice Unit, Netherlands Ministry of Education, Culture and Science, The Hague). Retrieved April 4, 2005, from http://www.eurydice.org/

Field, A. P. (2000). *Discovering statistics using SPSS for Windows: advanced techniques for the beginner*. London: Sage

Frederiksen, N. (1984). The real test bias, influences of testing on teaching and learning. *American Psychologist, 39*(3), 193-202.

Gibbs, G. (1992). *Improving the quality of student learning*. Bristol: Technical and Educational Services.

Gibbs, G. (1999). Using assessment strategically to change the way students learn. In S. Brown & A. Glasner (Eds.), *Assessment matters in higher education* (pp. 41-53). Buckingham: Open University Press.

Gielen, S., Dochy, F., & Dierick, S. (2003). Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including the pre-, post-, and true assessment effects. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of quality and standards*, (pp 37-54). Dordrecht: Kluwer Academic Publishers.

Gijbels, D. (2005). *Effects of new learning environments. Taking students' perceptions, approaches to learning and assessment into account*. Unpublished doctoral dissertation, University of Maastricht, The Netherlands.

Glaser, R., & Silver, E. (1993). Assessment, testing and instruction: Retrospect and prospect. *Review of Research in Education, 20*, 393-419.

Govaerts, M. J. B., Van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2005). The use of observational diaries in in-training evaluation: Student perceptions. *Advances in Health Sciences Education, 10*, 171-188.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Londen: Londen Sage.

Gulikers, J. T. M., Bastiaens, Th. J., & Martens, R. L. (2005). The surplus value of an authentic learning environment. *Computers in Human Behavior, 21*, 509-521.

Gulikers, J. T. J., Bastiaens, Th.J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67-85.

Gulikers, J. T. M., Bastiaens, Th. J., Kirschner, P. A., & Kester, L. (in press). Relations between student perceptions of assessment authenticity, study approach and learning outcome. *Studies in Educational Evaluation.*

Gulikers, J. T. M., Bastiaens, Th. J., Kirschner, P. A. (2006). *The practical value of the five-dimensional framework for assessment authenticity: student and teacher perceptions.* Manuscript submitted for publication.

Gulikers, J. T. M., Kester, L., Kirschner, P. A., & Bastiaens, Th. J. (2006). *Getting the whole picture: student, teachers and practitioner beliefs about authentic assessment.* Manuscript submitted for publication.

Hager.P., Gonczi, A., & Anthanasou, J. (1994). General issues about assessment of competence. *Assessment and Evaluation in Higher Education, 19*, 3-16.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5*, 1-16.

Handal, G., & Hofgaard Lycke, K. (2005, August). *From higher education to professional practice: learning among students and novice professionals*. Paper presented at the 11th biannual meeting of the European Association for Research on Learning and Instruction, Nicosia, Cyprus.

Hart, D. (1994). *Authentic Assessment: A Handbook for Education*. Menlo Park, CA: Addison-Wesley Publishing Company.

Henderson, P., & Karr-Kidwell, P. J. (1998). *Authentic assessment: An extensive literature review and recommandations for administrators*. (ERIC Document Reproduction Service No. ED418140)

Herrington, J., & Herrington, A. (1998). Authentic assessment and multimedia: how university students respond to a model of authentic assessment. *Higher Educational Research and Development, 17*(3), 305-322.

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational Technology Research and Development, 48*(3), 23-48.

Honebein, P. C., Duffy, T. M., & Fishman, B. J. (1993). Constructivism and the design of learning environments: Context and authentic activities for learning. In T. M. Duffy, J. Lowyck, & D. H. Jonassen (Eds.), *Desinging environments for constructive learning* (pp. 88-108). Berlin: Springer-Verslag.

Huang, H. M. (2002). Towards constructivism for adult learners in online learning environments. *British Journal of Educational Technology, 33*, 27-37.

Johnson, K. (1998). *Readability,* Retrieved December 6, 2004, from the Timetabler Web Site: http://www.timetabler.com

Joreskog K.G. (1993). Testing stuctural equation modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage Publishers.

Kalat, J. (1995). *Biological psychology*. Pacific Grove, CA: Brooks/Cole.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23-32.

Kasworm, C. E., & Marienau, C. A. (1997). Principles for assessment of adult learning. *New Directions of Adult and Continuing Education, 75*, 5-16.

Kerka, S. (1992). *Higher order thinking skills in vocational education* (ERIC Digest No. 127). (ERIC Document Reproduction Service No ED350487)

Kerka, S. (1995). Techniques for authentic assessment (Practical Application Brief). Columbus: ERIC Clearinghouse on Adult, Career, and Vocational Education, Center on Education and Training for Employement, The Ohio State University. (ERIC Document Reproduction Service No. ED381688)

Klare, G. R. (1963). *The Measurement of Readability*. Iowa: Iowa State University Press.

Klarus, R. (2000). Beoordeling en toetsing in het nieuwe onderwijsconcept [Evaluation and assessment in the new educational philosophy]. In J.Onstenk (Ed.), *Op zoek naar een krachtige beroepsgerichte leeromgeving. Fundamenten voor een onderwijsconcept voor de bve-sector*. 's Hertogenbosch: Cinop.

Klarus, R. (2003). *Instrument voor competentiegericht toetsen en beoordele [Instrument for competecy-based assessment and evaluation].* Retrieved December 6, 2003, from the Knowledge Centre for prior learning, assessment and recognition Web Site: http:/www.kenniscentrumevc.nl/site/documenten/KC3.pdf

Knowledge Centre for Vocational Education and Business (2004, April). Beroepscompetentieprofiel: verpleegkundige MBO [Professional competence profile: nursing VET]. Retrieved March 2, 2006, from the OVBD Web Site: http://www.ovdb.nl/

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15-21.

Lizzio, A., & Wilson, K. (2004a). First-year students' perceptions of capability. *Studies in Higher Education, 29*, 109-128.

Lizzio, A., & Wilson, K. (2004b). Action learning in higher education: an investigation of its potential to develop professional capability. *Studies in Higher Education, 29*, 469-488.

Lizzio, A., Wilson, K., & Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: implications for theory and practice. *Studies in Higher Education, 27*, 27-51.

Maclellan, E. (2001). Assessment for learning: the differing perceptions of tutors and students. *Assessment and Evaluation in Higher Education, 26*, 307-318.

Martens, R., Bastiaens, Th., & Gulikers, J. (2002). Leren met computergebaseerde authentieke taken: motivatie, gedrag en resultaten van studenten [Learning with computer-based authentic tasks: student motivation, behaviour and results]. *Pedagogische Studiën, 79*(6), 469-482.

Martens, R., Gulikers, J., & Bastiaens, Th. (2004). The impact of intrinsic motivation in e-learning with authentic computer tasks. *Journal of Computer Assisted Learning, 20*, 368-376.

Marton, F., & Säljö, R. (1997). Approaches to learning. In F. Marton, D. Hounsell, & N. Entwistle (Eds.), *The experience of learning. Implications for teaching and studying in higher education* (2nd. ed.) (pp. 39-59). Edinburgh: Scottish Academic Press.

McDowell, L. (1995). The impact of innovative assessment on student learning. *Innovations in Education and Training International, 32*, 302-313.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*, 13-23.

Meyer, C. (1992). What's the difference between authentic and performance assessment? *Educational Leadership, 49*(8), 39-40.

Miller, G. (1990). The assessment of clinical skills/competence/-performance. *Academic Medicine, 65*(9), 63-67.

Ministry of Education, Culture and Science (n.d.) Leren en werken: Plan van aanpak 2005-2007 [Learning and working: Action plan for 2005-2007]. Retrieved March 14, 2006, from http://www.minocw.nl

Morgan, D. L. (1997). *Focus groups as qualitative research*. London: Sage Publications.

Newmann, F. M., & Wehlage, G. G. (1993). Five standards for authentic instruction. *Educational Leadership, 50*(7), 8-12.

Newmann, F. M. (1997). Authentic assessment in social studies: Standards and examples. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp. 359-380). San Diego, CA: Academic Press.

Ngar-Fun, L. (2005). *Hong Kong academics' and students' perceptions of assessment purposes and practices*. The Hong Kong Institute of Education: Hong Kong.

Nicaise, M., Gibney, T., & Crane, M. (2000). Toward an understanding of authentic learning: student perceptions of an authentic classroom. *Journal of Science Education and Technology, 9*, 79-94.

Onstenk, J. (1997). *Lerend leren werken. Brede vakbekwaamheid en de integratie van leren, werken en innoveren* [Learning to learn-to-work: Broad job competencies and the integration of learning, working and innovating]. Delft: Eburon.

Orrell, J. (2003). *An exploration of congruence and disjunctions between academics' thinking when assessing and their beliefs about assessment practice.* Paper presented at the 11th Improving Student Learning Symposium: Research and Scholarship, Hinckley, England.

Pena, E. (1997). Great expectations: the reality of the workplace. *Australian Journal of Career Development, 6*, 25-32.

Perkins, D., & Blythe, T. (1994). Putting understanding up front. *Educational Leadership, 51*(5), 4-7.

Petraglia, J. (1998). *Reality by design: The rhetoric and technology of authenticity in education*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Prodromou, L. (1995). The backwash effect: from testing to teaching. *ELT Journal, 49*(1), 13-25.

Radinsky, J., Bouillion, L., Lento, E. M., & Gomes, L. M. (2001). Mutual benefit partnership: a curricular design for authenticity. *Journal of Curriculum Studies, 33*, 405-430.

Ramsden, P. (1992). *Learning to teach in higher education*. London: Routledge.

Reeves, T. C., & Okey, J. R. (1996). Alternative assessment for constructivist learning environments. In B. G. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 191-202). Englewood Cliffs, NJ: Educational Technology Publications.

Resnick, L. B. (1987). Learning in school and out. *Educational Leadership, 16*(9), 13-20.

Richardson J. T. E. (2005). Students' approaches to learning and teachers' approaches to teaching in higher education. *Educational Psychology, 25*, 673-680.

Roelofs, E., & Terwel, J. (1999). Constructivism and authentic pedagogy: state of the art and recent developments in the Dutch national curriculum in secondary education. *Journal of Curriculum Studies, 31*, 201-227.

Rust, C. (2002). The impact of assessment on student learning. *Active Learning in Higher Education, 3*, 145-158.

Sambell, K., McDowell, L., & Brown, S. (1997). But is it fair?: An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation, 23*, 349-371.

Sambell, K., & McDowell, L. (1998). The construction of the hidden curriculum: Messages and meanings in the assessment of student learning. *Assessment and Evaluation in Higher Education, 23*, 391-402.

Samuelowicz, K., & Bain, J. D. (1992). Conceptions of teaching held by academic teachers. *Higher Education, 24*, 93-111.

Samuelowicz, K., & Bain, J. D. (2002). Identifying academics' orientations to assessment practices. *Higher Education, 43*, 173-2001.

Savery, J. R., & Duffy, T. M. (1995). Problem based learning: An instructional model and its constructivist framework. In B. G. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design*. Englewood Cliffs, NJ: Educational Technology Publications.

Schnitzer, S. (1993). Designing an authentic asessment. *Educational Leadership, 50*(7), 32-35.

Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2004). Changing education, changing assessment, changing research? *Medical Education, 38*, 805-812.

Scouller, K. (1995). Different learning approaches of undergraduate students in two assessment contexts. *Research and Development in Higher Education, 17*, paper 89.

Scouller, K. (1997). Students' perceptions of three assessment methods: Assignment essay, multiple choice question examination, short answer examination. *Research and Development in Higher Education, 20*, 646-653.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice questions versus assignment essay. *Higher Education, 35*, 453-472.

Scouller, K., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Higher Education, 19*, 267-279.

Segers, M. S. R. (2004). Assessment en leren als een twee-eenheid: onderzoek naar de impact van assessment op leren [Assessment and learning as twofoldness: reserach on the impact of assessment on learning]. *Tijdschrift voor Hoger Onderwijs, 22(4)*, 188-220.

Segers, M., Dierick, S., & Dochy, F. (2001). Quality standards for new modes of assessment. An exploratory study of the consequential validity of the OverAll test. *European Journal of Psychology of Education, 16*(4), 569-586.

Segers, M., Dochy, F., & Cascallar, E. (2003). *Optimising new modes of assessment: In search of qualities and standards*. Dordrecht: Kluwer Academic Press.

Segers, M., Dochy, F., & De Corte, E. (1999). Assessment practices and students' knowledge profiles in a problem-based curriculum. *Learning Environments Research, 2*, 191-213.

Semeijn, J. (2005). *Academic competences and labour market entry. Studies among Dutch graduates*. Unpublished doctoral dissertation, University of Maastricht, The Netherlands.

Slavin, R. E. (1989). Research on cooperative learning: An international perspective. *Journal of Educational Research, 33*, 231-243.

Sluijsmans, D. (2002). *Student involvement in assessment: The training of peer assessment skills*. Unpublished doctoral dissertation, Open University of the Netherlands, The Netherlands

Smith, L., & Koshy, S. (2005, June). *Improving student learning through authentic assessment*. Paper presented at the First International Conference Enhancing Teaching and Learning through Assessment, Hong Kong.

Stein, S. J., Isaacs, G., & Andrews, T. (2004). Incorporating authentic learning experiences within a university course. *Studies in Higher Education, 29*, 239-258.

Sternberg, R. J. (1999). Cognitive Psychology (2nd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Straetmans, G. J. J. M., Sluijsmans, D., Bolhuis, B. G., & Van Merriënboer, J. J. G. (2003). Integratie van instructie en assessment in competentiegericht onderwijs [Integration of instruction and assessment in competency-based education]. *Tijdschrift voor Hoger Onderwijs, 3*, 171-197.

Struyven, K., Dochy, F., & Janssens, S. (2003). Students' perceptions about new modes of assessment in higher education: a review. In M. Segers, F. Dochy, & Cascallar E. (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 171-224). Dordrecht: Kluwer Academic Publishers.

Struyven, K. (2005). *The effects of student-activated teaching/learning environments on students' perceptions, student performance and pre-service teachers' teaching*. Unpublished doctoral dissertation, University of Leuven, Belgium.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.

Tang, C. (1994). Effects of modes of assessment on students' preparation strategies. In G. Gibbs (Ed.), *Improving student learning - Theory and practice* (pp. 151-170). Oxford: Oxford City Centre for Staff Development.

Thomas, P. R., & Bain, J. D. (1984). Contextual dependence of learning approaches: The effects of assessments. *Human Learning, 3*, 227-240.

Tigelaar, D. (2005). *Design and evaluation of a teaching portfolio*. Unpublished doctoral dissertation, University of Maastricht, The Netherlands.

Tillema, H. H., Kessels, J. W. M., & Meijers, F. (2000). Competencies as building blocks for integrating assessment with instruction in vocational education: a case from the Netherlands. *Assessment and Evaluation in Higher Education, 25*, 265-278.

Torrance, H. (1995). *Evaluating Authentic Assessment*. Buckingham: Open University Press.

Uhlenbeck, A (2002). *The development of an assessment procedure for beginning teachers of English as a foreign language.* Unpublished doctoral dissertation, University of Leiden, The Netherlands.

Van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: from methods to programmes. *Medical Education, 39*, 309-317.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2006). *Effects of sequencing process-oriented and product-oriented worked-out examples on troubleshooting transfer performance.* Manuscript submitted for publication.

Van Merriënboer, J. J. G. (1997). *Training complex cognitive skills: a four-component instructional design model for technical training*. Englewood Cliffs, NJ: Educational Technology Publishers.

Van Rossum, E. J., Deijkers, R., & Hamer, R. (1985). Students' learning conceptions and their interpretations of significant educational concepts. *Higher Education, 14*, 617-641.

Van Rossum, E. J., & Hamer, R. (2003, September). *Learning and teaching: a model of linked continua of conceptions.* Paper presented at the 11th Improving Student Learning Symposium: Research and Scholarship, Hinckley, England

Van Rossum, E. J., & Schenk, S. M. (1984). The relationship between learning conceptions, study strategies and learning outcome. *British Journal of Educational Psychology, 54*, 73-83.

Velde C. (1999). An alternative conception of competence: implications for vocational education. *Journal of Vocational Education and Training, 51*, 437-447.

Verhoeven, P., & Verloop, N. (2002). Identifying changes in teaching practice: innovative curricular objectives in classical language and the taught curriculum. *Journal of Curriculum Studies, 34*, 91-102.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46*(7), 41-47.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA, US: Jossey-Bass/Pfeiffer.

Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the course experience questionnaire. *Studies in Higher Education, 22*, 33-53.

Winning, T., Elaine, L., & Townsend, G. (2005). Student experiences of assessment in two problem-based dental curricula: Adelaide and Dublin. *Assessment and Evaluation in Higher Education, 30*, 489-505.

Wonacott, M. E. (2000). Vocational education research trends. *Trends and Issues Alert, 15.* Retrieved december 13, 2005, from http://www.calpro-online.org/eric/docs/tia00083.pdf

Summary

Authenticity is one of the most crucial elements of new modes of assessment (Dochy, 2001). Authentic assessments need to bridge the gap between what students have to do in assessments in school and what they, now or later, have to do in their internships or profession. Authentic assessments are expected to positively influence learning and stimulate students to develop professionally relevant competencies. These kinds of assessment are being developed in various types of education, but especially in Vocational Education and Training (VET), in which the main purpose is to prepare students for the workplace of today, authentic assessments are very relevant (Kerka, 1995). A problem, however, is that authenticity is not clearly defined and there are no clear guidelines for making assessments more authentic.

Additionally, authenticity is not a completely objective concept, rather it is partly subjective. This means that assessment authenticity depends on who is looking at it. This implies, among other things, that an assessment that is authentic in the eyes of a teacher, is not necessarily authentic in the eyes of a student. However, *student perceptions* of assessment characteristics strongly determine student learning (e.g., Scouller, 1998). It is expected that when a student perceives an assessment as more authentic, this positively influences his/her learning and competency development.

It is, thus, very relevant to examine when students perceive an assessment as being authentic and whether or not this actually stimulates their learning and competency development. In this thesis it is hypothesised that the perception of assessment authenticity depends on (1) a number of characteristics *in* an assessment and (2) the frame of reference (i.e., beliefs) of a person that is built on previous experiences with working in practice and/or experiences with assessments, authentic or not. The research questions that are addressed in this thesis are:

1. What characteristics of an assessment determine its (objective) authenticity?
2. Do these characteristics also determine authenticity in the eyes of students and teachers?
3. Do groups with different previous work and/or assessment experiences (i.e., students, teachers, and/or practitioners) have different beliefs about assessment authenticity? And does this influence how a person perceives and judges the authenticity of a new assessment?
4. What is the influence of student perceptions of assessment authenticity on their study approach and learning outcome?
5. Is the influence of student perceptions of assessment authenticity on their study approach and learning outcome dependent on the previous experiences of these students?

This thesis describes a number of studies, conducted in VET, in which groups with different previous experiences participate. These groups represent teachers, freshman and senior students, and practitioners. The ultimate goal is to determine guidelines for the development of authentic assessments that are authentic in the eyes of the student and stimulate the learning of students with different previous experiences.

Chapter 2 describes a literature review and an explorative study concerning the question of what assessment authenticity actually is. In other words, *what* or *which characteristics* determine the authenticity of an assessment? To define authenticity, we first needed to answer *in relation to*

what the authenticity of an assessment has to be judged (Messick, 1994). Since the goal of authentic assessment is to bridge the gap between learning in school and working in practice, authenticity is defined, in this thesis, as '"resembling professional practice". However, authenticity is a continuum. The *degree of assessment authenticity* depends on the degree of resemblance between the assessment and the professional practice situation the assessment aims to reflect. In addition, it is argued that several assessment characteristics influence the authenticity of an assessment. Thus, authenticity is not only a continuum, it is also multidimensional.

This led to the development of a five-dimensional framework (5DF) for assessment authenticity. This framework describes five characteristics that can resemble professional practice to a certain extent, thereby influencing the authenticity of the assessment. These five characteristics are: (1) the assessment task, (2) the physical context of the assessment (3) the social context of the assessment, (4) the result/form of the assessment, and (5) the assessment criteria. These five dimensions can be further divided into several characterising elements that contribute to the authenticity of the dimension. The 5DF is used as the starting point for examining and comparing authenticity beliefs and perceptions of various participant groups and to study their influence on student learning.

In an explorative, mostly qualitative, study the completeness of the 5DF was tested against the beliefs and perception of freshman students, senior students and teachers. At the same time, the relative importance of the five dimensions was examined. Additionally, differences and similarities between the participants were explored. The groups performed several individual and collaborative activities which led to the results that: (1) the 5DF seemed to be a good and fairly complete description of the characteristics that determine assessment authenticity with the marginal note that practitioners should be more involved, particularly in the development of authentic assessment criteria; (2) the task, result/form and criterion dimensions were found to be very relevant for authentic assessment, while the social context was perceived as the least important, and the opinions concerning the physical context were divided, and (3) senior students and teachers agreed more with each other as well as with the ideas of the 5DF, while freshman students seemed to have more traditional assessment beliefs that influenced how they perceived authenticity. This study seems to support, at least partly, the hypothesis that previous experiences influence a person's beliefs and perceptions of authenticity.

Chapter 3 examines whether the five dimensions and their characterising elements are recognised in practice as being important characteristics of authenticity. For this purpose, a perception questionnaire was developed based on the 5DF. Every scale intended to measure the resemblance between one dimension (e.g., the assessment task) and professional practice. Reliability analyses suggested that teachers supported the dimensions of authenticity as well as their characterising elements, while students only seemed to support three dimensions and one characterising element. Two possible explanations were explored to explain the results of the student, namely (1) students perceived authenticity differently than the 5DF presupposed, which would become visible in a different factor structure of a factor analysis, or (2) de scales of the

questionnaire exceeded the normally expected reading level of VET student, which made it difficult for student to understand the scales.

A factor analysis showed that students did indeed structure authenticity a bit differently that the 5DF presupposed. The two most important findings were that students seemed to perceive authenticity as being a multidimensional concept and that they seemed to recognise four of the five dimensions (except for the social context), while they did not identify the more specific characterising elements (except for criterion transparency). The *task* and *physical context* dimensions were recognised as such, while the students clustered the *form*-items in a separate factor and the *result* and *criterion*-items together. The new factors were reliable for both students and teachers. With respect to the second possible explanation, readability analyses showed that the unreliable scales exceeded the reading level of VET-students, while the reliable scales did not exceed the reading level of these students.

These results led to the conclusion that the 5DF distinguishes several determining dimensions of authenticity, because teachers, in any case, recognised all the dimensions and characterising elements, but that some restructuring is needed, based on the factors found in the student group. For educational practices these findings indicate that the 5DF offers concrete starting points for teachers to develop authentic assessments. However, when we want to influence the authenticity-perceptions of students, assessment should be changed at the dimensional level. Based on the factors found factors, a new questionnaire was developed at the dimensional level, which is used in the follow-up studies.

Chapter 4 describes the beliefs about assessment authenticity of different stakeholders, namely nursing students, teachers, and practitioners. These parties differed both in their work experiences as well as in their experiences with (authentic) assessments. The freshman students had relatively little, but current, experience with working in practice and they were familiar with authentic assessments at these workplaces. Through focus groups, individual interviews, and a questionnaire, we examined what the three groups believed to be important in an authentic assessment. The beliefs were examined, structured and compared by using the 5DF. Concerning the operationalisation of the *task* and the *physical context*, the three groups agreed. However, with respect to the *social context*, *form* and *criteria*, teachers differed in their beliefs compared to students and practitioners. The latter groups argued that teachers were not always up-to-date with the developments in the nursing profession. Moreover, they argued that teachers focused too much on technical skills at the expense of generic, professional skills. This study seems to support that previous experiences with respect to both working in practice and authentic assessment at the workplace influence beliefs about assessment authenticity.

Chapter 5 describes a study in which authenticity perceptions were compared between: (1) students and teachers and (2) freshman and senior students with different amounts of work experiences and different authentic assessment experiences (i.e., in and out of school). With the renewed perception questionnaire, all groups judged a role-play assessment on the five authenticity dimensions. The main results were that teachers perceived all the dimensions as being more authentic than students did, while freshman and senior students did not differ in how hey perceived the authenticity of the dimensions, contrary to the findings in chapter 2. This

might imply that teachers develop assessments according to their ideas of authenticity, which do not have to be in line with student perception of authenticity. Additionally, this study suggests that particularly *real* working experience (i.e., not during internships throughout schooling) changes perception of authenticity.

Chapter 6 and 7 both deal with the relationships between student perception of authenticity, study approach (deep or surface) and learning outcome (grades and/or generic skill development). In both studies it was hypothesised that an increased perception of authenticity of one or more of the dimensions of the 5DF leads to more deep learning and improved learning outcomes. These hypothesised relationships were modelled in a theoretical structural model. In chapter 6 this model is tested *within* one student group to find out if the hypothesised relationships exist and if they do, how they work (directly and/or indirectly). The expected relationships were for the most part supported. However, an unexpected finding was that an increased perception of criterion authenticity led to less deep learning and indirectly also to less development of generic skills. Qualitative data are needed to gain an insight into these negative relationships.

Chapter 7 tests the same theoretical model, but adds qualitative data and compares the variables and their relationships within the model *between* two student groups that differ in their amount of practical experience. Both groups judged the authenticity of a similar assessment (a role-play). Contrary to the findings in chapter 5, freshman students seemed to perceive the task and the physical context as being more authentic than senior students did. Freshman also reported more development of generic skills in response to this assessment. Moreover, most hypothesised relationships were supported in both student groups, corroborating the expectation that increased perceptions of assessment authenticity led to more deep learning and development of generic skills. However, one salient difference was found between both student groups. Experienced students showed a negative influence of perceived criterion authenticity on deep learning and development of generic skills, while this influence was, in line with the hypothesis, positive for freshman students. Qualitative data were used to find explanations for the found differences in the quantitative data. It was found that (1) senior students perceived the task as being too focused on general social work activities, (2) senior students argued that authentic assessment needs to be performed in the workplace, and (3) both freshman and senior students perceived the assessment criteria as being authentic, but senior students perceived the criteria as too analytic which inhibited them from performing the assessment task as they would do this in practice. Freshman students, on the other hand, said they needed these stepwise criteria to guide their learning. We concluded that analytic criteria subdivide performance in several authentic steps that adequately describe *how* the task should be performed in the eyes of freshman students. However, more experienced students do not perceive performing in practice in this stepwise manner anymore, because of their increased amount of experience with performing these kinds of tasks in real professional practice. Stepwise assessment criteria do not reflect *how* these students perceive performing in professional practice, which made these criteria not beneficial for the learning of more experienced students.

Chapter 8 combines the findings of the studies and reflects on the dimensions of the 5DF though the eyes of the different beholders. This reflection shows three main conclusions of this thesis, namely that assessment authenticity depends on (five) different assessment characteristics, that assessments, which students perceive as being authentic, are positive for learning, but that *what* students perceive as authentic depends on their previous work and/or assessment experiences. This means that an optimal operationalisation of authentic assessments differs for students with more or less work and/or assessment experiences. A number of practical implications and guidelines are described for the development of authentic assessment for students with different previous experiences.

Samenvatting

Authenticiteit is een van de belangrijkste kenmerken van nieuwe vormen van assessment (Dochy, 2001). Authentieke assessments[8] moeten een brug slaan tussen wat studenten moeten doen in toetsen op school en wat ze, nu of later, moeten doen in hun stage of beroep. Van authentieke assessments wordt verwacht dat ze een positieve uitwerking hebben op leergedrag en studenten stimuleren om beroepsrelevante competenties te ontwikkelen. In verschillende onderwijscontexten worden authentieke assessments ontwikkeld, maar vooral in het Middelbaar BeroepsOnderwijs (MBO), waar het belangrijkste doel is studenten voor te bereiden op de beroepspraktijk, is authentieke assessment zeer relevant (Kerka, 1995). Een probleem is echter dat authenticiteit nog niet duidelijk gedefinieerd is en dat er geen duidelijke richtlijnen zijn voor hoe een assessment daadwerkelijk meer authentiek kan worden gemaakt.

Daarnaast is authenticiteit geen volledig objectief concept, maar is het deels subjectief. Dat betekent dat assessment authenticiteit wordt bepaald door degene die dit assessment ervaart. Dit impliceert onder andere dat een assessment dat authentiek is in de ogen van een docent, niet authentiek hoeft te zijn in de ogen van de student. Echter, *studentpercepties* van assessment kenmerken zijn bepalend voor het leergedrag (b.v. Scouller, 1998). Verwacht wordt dat wanneer een student een assessment als meer authentiek percipieert, dit een positieve invloed heeft op zijn/haar leren en competentieontwikkeling.

Het is dus zeer relevant om te onderzoeken wanneer een student een assessment als authentiek percipieert en of dit ook daadwerkelijk een positieve bijdrage levert aan zijn/haar leren en competentieontwikkeling. In dit proefschrift wordt verondersteld dat de perceptie van assessment authenticiteit bepaald wordt door (1) een aantal kenmerken *in* een assessment en (2) het referentiekader (beliefs[9]) van een persoon, dat ontstaat op basis van eerdere werkervaringen en/of ervaringen met (authentieke) assessments. De onderzoeksvragen die in dit proefschrift aan bod komen, zijn:

1. Welke kenmerken van een assessment bepalen de (objectieve) authenticiteit?
2. Zijn deze kenmerken ook bepalend voor de authenticiteit in de ogen van studenten en docenten?
3. Hebben groepen met verschillende eerdere werk en/of assessment ervaringen (studenten, docenten en het werkveld) andere referentiekaders ten aanzien van assessment authenticiteit? En beïnvloedt dit hoe iemand de authenticiteit van een nieuw assessment bekijkt?

---

[8] Er is geen goede Nederlandse vertaling van het woord assessment. Twee Nederlandse woorden die hierbij in de buurt komen zijn, beoordelen of beoordelingvormen, in tegenstelling tot het meer traditionele toetsen. Er is voor gekozen om ook in de Nederlandse samenvatting te spreken van assessment

[9] Er is geen goede Nederlandse vertaling voor het woord beliefs. Het woord dat erbij in de buurt komt is overtuiging. Er is voor gekozen om ook in de Nederlandse samenvatting te spreken van beliefs

4. Wat is de invloed van studentpercepties van assessment authenticiteit op studeergedrag en leeruitkomst?
5. Is de invloed van studentpercepties van assessment authenticiteit op hun studeergedrag en leeruitkomst afhankelijk van eerdere ervaringen van studenten?

Dit proefschrift beschrijft een aantal studies, uitgevoerd in het MBO, waarin groepen met verschillende eerdere ervaringen aan bod komen, te weten docenten, beginnende en meer ervaren studenten en mensen uit het werkveld. Het uiteindelijke doel is om richtlijnen te kunnen geven voor het ontwikkelen van authentieke assessments die authentiek zijn in de ogen van de student en bijdragen aan het leren van studenten met verschillende soorten ervaringen.

Hoofdstuk 2 beschrijft een literatuurstudie en een exploratieve studie naar de vraag wat assessment authenticiteit is. Met andere woorden, *wat* of *welke kenmerken* bepalen de authenticiteit van een assessment? Alvorens een definitie van authenticiteit kan worden gegeven, moet eerst bepaald zijn *ten opzichte van wat* de authenticiteit van een assessment moet worden beoordeeld (Messick, 1994). Omdat het doel van authentieke assessments is een brug te slaan tussen leren op school en werken in de praktijk, wordt authenticiteit in dit proefschrift gedefinieerd als "overeenkomstig de beroepspraktijk". Echter, authenticiteit is een continuüm. De *mate van* assessment authenticiteit wordt bepaald door de mate van overeenkomst tussen het assessment en de beroepssituatie die dit assessment weerspiegelt. Bovendien wordt beargumenteerd dat authenticiteit bepaald wordt door verschillende assessment kenmerken. Met andere woorden, authenticiteit is niet alleen een continuüm, het is tevens multidimensionaal.

Dit heeft geleid tot de ontwikkeling van een vijf-dimensionaal model (5DM) voor assessment authenticiteit. Dit model beschrijft vijf kenmerken die in meer of mindere mate overeenkomen met de beroepspraktijk en daarmee de authenticiteit van het assessment beïnvloeden. Deze vijf kenmerken zijn: (1) de assessment taak, (2) de fysieke context waarin het assessment plaatsvindt, (3) de sociale context van het assessment, (4) de resultaat/vorm van het assessment, en (5) de assessment criteria. In het 5DM worden deze dimensies verder opgesplitst in een aantal kenmerkende elementen dat bijdraagt aan de authenticiteit van de betreffende dimensie. Het 5DM wordt als uitgangspunt gebruikt voor het bestuderen en vergelijken van authenticiteitsbeliefs en -percepties van verschillende groepen deelnemers en de invloed hiervan op het leergedrag van studenten.

In een exploratieve, voornamelijk kwalitatieve, studie werd de volledigheid van het 5DM getoetst aan de beliefs en percepties van beginnende studentgroepen, meer ervaren studentgroepen en docenten. Tevens werd onderzocht of de vijf dimensies verschilden in belangrijkheid volgens de verschillende groepen. Bovendien werd gekeken naar verschillen en overeenkomsten *tussen* de groepen. Alle groepen voerden een aantal individuele en groepsactiviteiten uit met als belangrijkste resultaten: (1) Het 5DM leek een goede en vrijwel volledige beschrijving te geven van de kenmerken die belangrijk zijn voor authenticiteit, met de kanttekening dat het werkveld meer betrokken moest worden, in het bijzonder bij het ontwikkelen van authentieke beoordelingscriteria; (2) de taak, vorm/resultaat en criteria dimensies waren zeer belangrijk voor authenticiteit, terwijl de sociale context het minst

belangrijk leek en de meningen met betrekking tot de fysieke context sterk verschilden, en (3) ervaren studenten en docenten kwamen zowel met elkaar als met de ideeën van het 5DM overeen, terwijl beginnende studenten meer traditionele assessment beliefs leken te hebben, die bepaalden hoe ze tegen authenticiteit aankeken. Deze studie lijkt de verwachting dat eerdere ervaringen beliefs en percepties van authenticiteit beïnvloeden gedeeltelijk te ondersteunen.

In hoofdstuk 3 wordt op een kwantitatieve manier onderzocht of de vijf dimensies en de bijbehorende elementen worden herkend in de praktijk als zijnde kenmerkende elementen van authenticiteit. Voor dit doel werd een perceptievragenlijst ontwikkeld met schalen gebaseerd op het 5DM. Iedere schaal was gericht op de overeenkomst tussen één kenmerk (bv. de assessment taak) en de beroepspraktijk. Betrouwbaarheidsanalyses toonden aan dat docenten de dimensies alsmede de kenmerkende elementen van authenticiteit ondersteunden, terwijl studenten slechts drie dimensies en een bijbehorend element leken te ondersteunen. Twee mogelijke verklaringen werden nader onderzocht, namelijk: (1) Studenten percipieerden authenticiteit anders dan het 5DM veronderstelde, wat zichtbaar zou worden in een andere factor structuur in een factoranalyse, of (2) de schalen van de vragenlijst overschreden het normale leesniveau van de studenten, waardoor de schalen moeilijk te begrijpen waren voor studenten.

Een factoranalyse liet zien dat studenten het concept authenticiteit inderdaad enigszins anders structureerden dan het 5DM voorstelde. De twee belangrijkste bevindingen waren dat studenten authenticiteit wel als een multidimensionaal concept leken te zien en dat ze vier van de vijf dimensies leken te herkennen (behalve de sociale context), terwijl ze de meer specifieke kenmerkende elementen niet identificeerden (behalve criterium transparantie). De *taak* en de *fysieke context* kwamen overeen met het 5DM, echter studenten clusterden de *vorm* items in en aparte factor en de *resultaat* en *criteria* items samen. De ontstane factoren waren betrouwbaar voor zowel studenten als docenten. Met betrekking tot de tweede mogelijke verklaring lieten leesbaarheidsanalyses zien dat de onbetrouwbare schalen het normale leesniveau van MBO studenten overschreden, terwijl de betrouwbare schalen een normaal leesniveau van deze studenten vereisten.

Geconcludeerd wordt dat het 5DM verschillende belangrijke kenmerken van authenticiteit onderscheidt, omdat alle kenmerken in ieder geval door docenten werden herkend, maar dat enige herstructurering nodig is gebaseerd op de gevonden factoren in de studentgroep. Voor de onderwijspraktijk impliceert deze studie dat het 5DM concrete aanknopingspunten voor docenten biedt om authentieke assessments te ontwikkelen. Echter, wanneer men de authenticiteitsperceptie van studenten wil beïnvloeden, is het belangrijk om assessments op dimensieniveau te veranderen. Op basis van de gevonden factoren werd een nieuwe vragenlijst gemaakt op dimensieniveau, die gebruikt wordt in de vervolgstudies.

Hoofdstuk 4 beschrijft de beliefs ten aanzien van assessment authenticiteit van verschillende partijen, namelijk studenten en docenten verpleegkunde en verpleegkundigen werkzaam in de praktijk. Deze drie partijen verschilden zowel in hun werkervaringen als in hun ervaringen met (authentieke) assessments. De eerstejaars studenten hadden relatief weinig, maar recente, ervaring met werken in de praktijk, maar hadden wel ervaring met authentieke assessment op deze werkplek. Door middel van focusgroepen, individuele interviews en een vragenlijst werd

onderzocht hoe deze drie groepen vonden dat een authentiek assessment eruit zou moeten zien. Deze beliefs werden onderzocht, gestructureerd en met elkaar vergeleken aan de hand van het 5DM. Over de vormgeving van de *taak* en de *fysieke context* waren de drie groepen het eens. Echter, ten aanzien van de *sociale context*, *vorm* en *criteria* hadden docenten andere beliefs dan studenten en verpleegkundigen. De laatste twee groepen vonden dat docenten vaak geen goed beeld hadden van de huidige beroepspraktijk. Tevens vonden zij docenten teveel gericht op technische vaardigheden ten koste van meer generieke, beroepsgerichte vaardigheden. Deze studie lijkt te bevestigen dat zowel werkervaringen als ervaringen met assessments op de werkplek invloed hebben op hoe verschillende partijen over assessment authenticiteit denken.

Hoofdstuk 5 beschrijft een studie waarin de authenticiteitspercepties werden vergeleken tussen: (1) studenten en docenten en (2) beginnende en ervaren studenten, die verschilden in zowel hoeveelheid werkervaring als ervaring met authentieke assessment op school en in de praktijk. Aan de hand van de vernieuwde perceptievragenlijst beoordeelden alle groepen een rollenspel assessment op de vijf authenticitieitsdimensies. De belangrijkste resultaten waren dat docenten alle dimensies authentieker vonden dan de studenten en dat de beginnende en ervaren studenten niet verschilden in hoe authentiek ze de dimensies vonden, in tegenstelling tot de resultaten uit de studie in hoofdstuk 2. Dit zou kunnen impliceren dat docenten toetsen ontwikkelen volgens hun ideeën van authenticiteit, die niet automatisch overeenkomen met wat studenten authentiek vinden. Bovendien suggereert deze studie dat vooral *echte* werkervaring (dus niet tijdens stages gedurende de opleiding) de perceptie van authenticiteit verandert.

Hoofdstuk 6 en 7 gaan beide in op de relaties tussen studentpercepties van authenticiteit, studeeraanpak (diep en oppervlakkig) en leeruitkomst (cijfer en/of ontwikkeling van generieke vaardigheden). In beide studies werd verwacht dat een stijging in authenticiteitsperceptie van (een of meer van) de vijf dimensies zou leiden tot meer diep leren en betere leeruitkomsten. Deze verwachte relaties werden gemodelleerd in een theoretisch structureel model. In hoofdstuk 6 wordt dit model onderzocht *binnen* een groep om uit te vinden of de verwachte relaties bestaan en zo ja, hoe de relaties lopen (direct en/of indirect). De verwachte relaties werden veelal ondersteund. Echter, een onverwachte bevinding was dat een stijging in authenticiteitsperceptie van de beoordelingscriteria een negatieve invloed had op diep leren en daarmee ook indirect op de gerapporteerde ontwikkeling van generieke vaardigheden. Kwalitatieve data leken nodig om meer inzicht te krijgen in de reden van dit negatieve verband.

Hoofdstuk 7 toetst hetzelfde theoretische model, maar voegt kwalitatieve data toe en vergelijkt bovendien de variabelen en hun relaties *tussen* twee studentgroepen die verschillen in hun mate van praktijkervaring. Beide groepen beoordeelden de authenticiteit van een soortgelijk assessment (een rollenspel). In tegenstelling tot de bevindingen in hoofdstuk 5, vonden beginnende studenten de *taak* en de *fysieke context* meer authentiek dan meer ervaren studenten. De beginners rapporteerden bovendien meer ontwikkeling van generieke vaardigheden. Verder ondersteunden beide groepen bijna alle verwachte relaties in het theoretische model. Hiermee werd de hypothese dat een stijging van authenticiteitperceptie positief is voor leren en ontwikkeling van generieke vaardigheden, voor een groot deel bevestigd. Er was echter één belangrijk verschil tussen beide groepen. Voor ervaren studenten was er een negatieve invloed

van authenticiteitsperceptie van de beoordelingscriteria op diep leren en ontwikkeling van generieke vaardigheden, terwijl deze invloed voor beginners, in lijn met de verwachting, positief was. Kwalitatieve data werden gebruikt om verklaringen te vinden voor de gevonden verschillen in de kwantitatieve data. Hieruit bleek dat (1) ervaren studenten de taak teveel gericht vonden op algemene sociaalwerk activiteiten, (2) ervaren studenten vonden dat authentiek toetsen op de werkplek moest plaatsvinden en (3) zowel beginnende als ervaren studenten de criteria authentiek vonden, maar dat ervaren studenten deze criteria te analytisch vonden waardoor ze de assessment taak niet konden uitvoeren zoals ze dat in de echte praktijk zouden doen. Beginners, daarentegen, zeiden deze stapsgewijze criteria nodig te hebben om hun leren te sturen. Geconcludeerd wordt dat analytische criteria taakuitvoering opsplitsen in verscheidene authentieke stappen die voor beginnende studenten een adequaat beeld geven van *hoe* deze taak in de praktijk uitgevoerd wordt. Echter, meer ervaren studenten zien, door hun ervaring met het uitvoeren van soortgelijke taken in de beroepspraktijk, taakuitvoering in de praktijk niet meer zo stapsgewijs. Stapsgewijze assessmentcriteria komen niet overeen met *hoe* deze studenten functioneren in de beroepspraktijk zien, waardoor deze criteria niet bijdragen aan het leren van meer ervaren studenten.

Hoofdstuk 8 combineert de bevindingen van alle studies en reflecteert op de vijf dimensies van het 5DM vanuit de visies van de verschillende groepen. Deze reflectie vertoont de drie belangrijkste conclusies van dit proefschrift, namelijk dat assessment authenticiteit afhankelijk is van (vijf) verschillende assessment kenmerken; dat assessments, die door studenten als authentiek gepercipieerd worden, positief zijn voor leren; maar dat *wat* studenten authentiek vinden afhankelijk is van hun werk en/of assessment ervaringen. Dit betekent dat een optimale vormgeving van authentieke assessments enigszins anders is voor studenten met meer of minder werk en/of assessment ervaring. Een aantal praktische implicaties en richtlijnen voor de ontwikkeling van authentieke assessment voor verschillende studentgroepen worden besproken.

Curriculum Vitae

Judith Gulikers was born 5 November 1979 in Maastricht, The Netherlands. She studied Cognitive/Educational Psychology at the University of Maastricht and followed a Work and Organisational Psychology major at the Tilburg University. Her master thesis, conducted at the Educational Technology Expertise Centre of the Open University of the Netherlands, concerned learning with authentic, computer-based tasks. After graduating in 2002, she started as a PhD-student on authentic assessment at the Educational Technology Expertise Centre of the Open University of the Netherlands. Currently, she is working at the Education and Competence Study Group of the Wageningen University, The Netherlands. Her main areas of research focus on the effects of competency-based assessments on student learning and competency-development in vocational types of education.

Main Publications

Scientific Publications

Gulikers, J. T. M., Bastiaens, Th. J., & Martens, R. L. (2005). The surplus value of an authentic learning environment. *Computers in Human Behavior, 21,* 509-521.

Gulikers, J. T. M., Bastiaens, Th. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research & Development, 52,* 67-85.

Martens, R., Gulikers, J., & Bastiaens, Th. (2004). The impact of intrinsic motivation on e-learning in authentic computer tasks. *Journal of Computer Assisted Learning, 20,* 368-376

Martens, R., Bastiaens, Th., & Gulikers, J. (2002). Leren met computergebaseerde authentieke taken: motivatie, gedrag en resultaten van studenten. *Pedagogische Studiën, 6,* 469-481.

Gulikers, J. T. M., Bastiaens, Th. J., & Kirschner, P. A. (2006). Authentic assessment, student and teacher perceptions: the practical value of the five-dimensional framework. *Journal of Vocational Education and Training, 58,* 337-357.

Gulikers, J. T. M., Bastiaens, Th. J., Kirschner, P. A., & Kester, L. (in press). Relations between student perceptions of assessment authenticity, study approaches and learning outcome. *Studies in Educational Evaluation.*

Book Chapters and Other Publications

Gulikers, J.T.M., Bastiaens, Th. J., & Kirschner, P.A. (in press). Defining authentic assessment : Five dimensions of authenticity. A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education.* New York : Routledge.

Gulikers, J., Bastiaens, Th., & Kirschner, P. (2005). Authentieke toetsing, de beroepspraktijk in het vizier [Authentic assessment, professional practice within sight]. *OnderwijsInnovatie, 2,* 17-24.

Gulikers, J. (2005). Ontwerpen van authentieke toetsen met de beroepspraktijk als uitgangspunt [Developing authentic assesssment with professional practice as starting point]. Handboek Effectief Opleiden, *39,* 33-53.